October 2021

Measurement Invariance Across Immigrant and Non-Immigrant Populations on PISA Cognitive and Non-Cognitive Scales

Maritza Casas
University of Massachusetts Amherst

Measurement Invariance Across Immigrant and Non-Immigrant Populations on PISA Cognitive and Non-Cognitive Scales

A Dissertation Presented

by

MARITZA CASAS

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September 2021

College of Education

Research in Educational Measurement and Psychometrics

Measurement Invariance Across Immigrant and Non-Immigrant Populations on PISA Cognitive and Non-Cognitive Scales

A Dissertation Presented

By

MARITZA CASAS

Approved as to style and content by:

_____
Stephen G. Sireci, Chair

_____
Craig S. Wells, Member

_____
Maria Elena Oliveri, Member

_____
Ezekiel Kimball
Associate Dean of Academic Affairs
College of Education

ACKNOWLEDGMENTS

The completion of this dissertation would have never been possible without the support of a lot of people. First, I want to thank God for being with me all the way. I also thank all the members of my family especially my sweet and loving GRANDMOTHER who has been with me every step of the way, my mom, sister, Adriana, Memo, Gladys, and Glory.

I also want to thank my first mentor, Alba Lucia Meneses who not only helped me become the professional and researcher that I am today but pushed me with her love and support to pursue my dreams. In the same line I want to thank Conny Del Portillo and her sweet family Marta Claudia and Luis Enrique who have been more than a real family to me in this country. Thank you for your genuine love and support and for always being there for me through the good and bad times.

I also want to thank Aura Nidia Herrera for admitting me into the Master's Program, for all the learning opportunities she provided me with, for her support, and for pushing me to come to Amherst, without her I would have never known UMass.

I want to thank Steve Sireci for welcoming me into the program when I first came to do an internship and for encouraging me to apply for the PhD. Thank you for opening the doors for me, for all the support you gave me since day one and throughout all these years, and for being my advisor.

I want to give thanks to Craig Wells for all the support he provided me especially while I was working on my dissertation. Thank you for your patience, willingness to help me, your time, and most importantly for believing in me and reminding me that I was not alone. This dissertation would have never been possible without your help and support.

And I also thank all the REMP family especially Duy Pham and my great cohort Gabriel (Joan and Sary), Jane, and Darius. I am glad to have met you all and most importantly to still have you in my life.

I want to give a special thank you to Ann Marie Russell and Barb Chalfonte. Without their support I would have never been able to complete my studies, find a job, and survive the pandemic, they are truly role models not only as amazing professionals but also as amazing human beings. I am very fortunate to have met you.

I also want to thank Karen and Joe for their constant love and support but more importantly for welcoming into their family since the first time I came to Amherst. You have been a real family to me and you are one of the best things that happened to me when I came to Amherst and I feel very lucky to have you in my life. I love you very much.

I finally want to thank all my the wonderful people who were an essential part of this journey: hermana Violeta, Maria Isabel Bulla, Catheryne Lancheros, Sandra Camargo (and Victor), Maria Elena Thalliens, Stefany Flores, Jenny Cardenas, Jazmine Escobar, Rocio Barajas, Diana Rodriguez, Yvonne Gomez, Yaneth Sanabria, Ha and Canh, Ming Coler, Sandra Sanchez, Nubia Castiblanco, Melba Ruiz, Maria Consuelo Leon, Joan Daniels, Padre Paolo, Diana Senior, Nelly Ayala, Madeleine Barrera, Hermana Susana, Hazel, Emily Stone, Father Gary, Father Rob, Father Francis, Ana Ortiz, Claudia Paez, Soeun Kim, Rajshree Pandey, Rosa Medina, and Arlen Marielos.

This has been quite a journey and when I look back I can only see all the wonderful people who stood by me and contributed to this achievement. Thank you all from the bottom of my heart!

ABSTRACT

MEASUREMENT INVARIANCE ACROSS IMMIGRANT AND NON-IMMIGRANT POPULATIONS ON PISA COGNITIVE AND NON-COGNITIVE SCALES

SEPTEMBER 2021

MARITZA CASAS, B.A., UNIVERSIDAD CATOLICA DE COLOMBIA

M.S., UNIVERSIDAD DE LOS ANDES

M.S., UNIVERSIDAD NACIONAL DE COLOMBIA

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Dr. Stephen G. Sireci

International large-scale educational assessments (ILSAs) have played a relevant role in educational policies targeting immigrant students across countries as their results are used by governments as input for decision-making purposes. Given the potential impact that ILSAs can have, the psychometric features of these assessments must be carefully assessed and empirical evidence about the extent to which the inferences made based on test results are valid must be collected. To do so, the first step is to determine if the test results have the same meaning across countries and groups of examinees that is, if the measures are invariant so that results can be compared directly among countries.

The general purpose of this dissertation was to provide evidence about the extent to which the 2018 Programme for International Student Assessment (PISA) provides invariant measures of reading literacy, exposure to bullying, and sense of belonging at school for immigrant students from diverse cultural and linguistic backgrounds across the countries that host large populations of immigrants. Moreover, given that test performance can be impacted by non-cognitive variables, the constructs exposure to

bullying and sense of belonging at school were analyzed as potential predictors of student performance in reading literacy.

Two modeling approaches were implemented to evaluate measurement invariance: a traditional approach (multiple group confirmatory factor analysis) and a more contemporary approach that has shown to be more suitable to handle the complex features of ILSAs. The overall results showed that the alignment optimization procedure was a more suitable statistical tool than the traditional modeling technique -multiple-group confirmatory factor analysis- for the evaluation of measurement invariance when the data under analysis are collected through ILSAs since it can handle the features and complexities of these data while allowing for the incorporation of the immigration status into the analysis.

The implications of the overall findings for educational policymakers, educators, test developers, and educational researchers were discussed along with five limitations that should be addressed in future studies.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

CHAPTER

I INTRODUCTION

1.1 International Migration

In the past decades, mass migration has significantly increased, especially in western industrialized or developed nations due to political instability, war, and economic catastrophes. According to migration trends, 35 to 40 million people migrate every five years. For instance, it has been estimated that the United States receives around 70,000 refugees and immigrants each year -mostly from Latin America and Asia- and in some European nations, the number of immigrants is even higher (Global Migration Data Analysis Centre & International Organization for Migration, 2018; Powers & Pivovarova, 2017; Rubinstein-Avila, 2016). Specifically, by 2017, the countries that hosted the largest number of international migrants included United States of America (50%), followed by Saudi Arabia (12%), Germany (12%), Russian Federation (11.7%), and United Kingdom (8.8%) (United Nations, 2017).

These large migration movements have promoted discussions about diversity at the social and policy levels among the host countries where the main challenge is to promote tolerance of diversity and ensure equitable participation within multicultural societies (Isac et al., 2019). As a result, most governments in the host countries have prioritized international migration in their political agendas in terms of the design of policies that promote the integration of immigrants into the social and economic systems while providing them with basic services and resources (Teltemann & Schunck, 2016; Volante et al., 2017).

1.2 Integration of Immigrants

Currently, host countries face the challenge of preserving social cohesion in the presence of these large migration flows through the integration of immigrants into the social systems so that they can acquire the necessary skills that will allow them to eventually join the labor market and make contributions to the overall economy as well as to the development of science and technology. In this sense, the education of this population is of great relevance (Borgonovi, 2018; Bozick et al., 2016; Global Migration Data Analysis Centre & International Organization for Migration, 2018; United Nations, 2017).

Educational systems are, therefore, a key element to effectively address diversity and integrate immigrants through the implementation of policies that facilitate the acquisition of educational skills and knowledge that are needed so that immigrant populations can have an active participation in the societies of the host countries while making both economic and sociocultural contributions (Borgonovi, 2018; Chiu et al., 2012; Rubinstein-Avila, 2016; Volante et al., 2017).

In fact, the effectiveness of educational systems in these international scenarios lies in their ability to successfully combine high levels of achievement with high levels of equity so that all students are provided with access to high-quality education and with the same opportunities to reach their full academic potential regardless of their cultural and ethnical backgrounds, economic status or personal circumstances (Organisation for Economic Co-operation and Development [OECD], 2016). To do so, efforts towards a better understanding of academic achievement among immigrant students are needed to enhance the theoretical understanding of the association between immigration and

education which could in turn guide the development of educational programs and more cohesive policy approaches to help schools meet the unique educational needs of the growing population of immigrant students providing them with the resources they need to succeed and become socially active and competent adults (Crosnoe & Turley, 2011; Duong et al., 2016; Pivovarova & Powers, 2019; Rubinstein-Avila, 2016).

Ideally, educational policies that effectively promote equity would have (a) rigorous and consistent educational standards for all the students, (b) pre-defined educational strategies targeting at-risk students, and (c) an equitable distribution of resources across schools (Organisation for Economic Co-operation and Development, 2016).

However, and despite the potential benefits that have been associated to migration, immigrants remain among the most vulnerable members of the society mostly due to a lack of well-managed migration and integration policies (United Nations, 2017). Moreover, the major weakness in integration policies throughout the world is education because there is typically little to no support for immigrant students who usually require additional educational, financial and social resources (Volante et al., 2017).

As a result, it is common to find gaps in terms of educational outcomes between immigrant and native students where the former tends to show lower academic performance. The differences between these groups of students in terms of socio-economic disadvantages, language proficiency, ethnicity, cultural background, and educational level of the parents, impact their academic performance and result in academic disparities (Borgonovi, 2018; Crosnoe & Turley, 2011; Giannelli & Rapallini, 2016).

Due to the lack of proper policies and strategies to handle these gaps, host nations have additionally reported to be faced with the challenge of managing high dropout rates among immigrant students and even though the overall proportion of immigrant students within schools is not high, they do account for a high percentage of the dropout population. These dropout rates have also been associated to absenteeism, poor school engagement, lack of sense of belonging at school, and work/family responsibilities (Rubinstein-Avila, 2016).

1.3 Educational Integration of Immigrants

In this general scenario, schools could play a key role in the integration of immigrant students by (a) promoting an active participation of immigrant students into the social lives of the communities where they belong, (b) contributing to their psychosocial wellbeing, (c) providing them with equal access to the academic curriculum to help them achieve their educational goals, and (d) providing them with the proper resources to meet their particular educational needs (Callahan et al., 2010; Chiu et al., 2012; Volante et al., 2017). However, the interplay among immigration, education and social mobility is so complex that immigrant students typically end up facing discrimination and other barriers in school that reinforce social stratification (Crosnoe & Turley, 2011).

Given the complexity and variety of factors that shape the experiences and educational outcomes of immigrant students at schools, the first step towards their proper integration into the educational systems would be to obtain accurate information about the current level of their academic skills and other individual variables that could potentially impact their academic performance so that schools can set up the basis for the

design of policies and programs that are tailored to their needs and facilitate their efficient integration into the school systems. To this end standardized educational assessments, especially international large-scale assessments (ILSAs), become a valuable source of information.

1.3.1 International Large-Scale Assessments

ILSAs collect data about educational systems throughout the world to inform educational policies, compare student achievement, support curriculum implementation, and inform educational decision-making processes. Specifically, ILSAs aim to compare academic proficiency across countries and identify the distribution of competencies at various educational stages to help countries detect issues within their national educational systems with respect to evidence-based criteria while providing a basis to judge their overall quality. Currently, more than 50% of the world's countries are taking part in these evaluations and the most used ILSAs are the Programme for International Student Assessment (PISA) and the Trends in International Mathematics Study (TIMSS). In fact, results from PISA are currently being used by governments throughout the world to improve educational policies and practices by setting policy targets against measurable goals achieved by other educational systems (Cordero et al., 2018; Crosnoe & Turley, 2011; Hopfenbeck et al., 2018; Kim et al., 2017; Oliveri et al., 2018; Oliveri & Ercikan, 2011; Oliveri & von Davier, 2011; OECD, 2016; van de Vijver et al., 2019).

Hopfenbeck et al. (2018) conducted a systematic literature review of published papers in peer-reviewed journals that reported research on PISA from 1999 to 2015 and found a significant increase in the number of publications reporting analysis from PISA from one article in 1999 to around 100 papers in 2015. The authors also found that most

articles published around the world focused on the evaluation of educational inequalities related to socio-economic status and migration, which reflects a major international concern to find better educational approaches that guarantee the effective educational inclusion of immigrant students into the educational systems.

Data from ILSAs are therefore increasingly being used not only as input for the formulation of educational policies, but also to compare educational outcomes within and between countries in an effort to identify key factors that contribute to the overall improvement of educational systems. However, in order to make those comparisons, test scores must be "comparable" to guarantee that the same target latent construct is being measured in the same way and holds the same meaning across all the countries (Byrne et al., 2009; Crosnoe & Turley, 2011; Hox et al., 2012; Isac et al., 2019; Kim et al., 2017; Oliveri & Lawless, 2018; Oliveri & von Davier, 2011; Oliveri & von Davier, 2016; Sireci, 2015; van de Vijver et al., 2019; Volante et al., 2017). Specifically, score comparability refers to the extent to which test scores have a consistent meaning across cultural groups (Oliveri et al., 2015).

Additionally, when ILSAs are administered to highly diverse populations, actions must be implemented to ensure that all test takers will be assessed fairly regardless of their differences in characteristics that are not relevant for the measurement of the target latent construct (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 2014; Educational Testing Service, 2014). Several threats to the validity of interpretations made from ILSAs have been identified: (a) method bias, a systematic source of error that impacts scores due to bias associated to the measurement process (e.g., administrator

bias, familiarity with item formats); (b) construct bias that refers to inconsistency of the target latent construct across groups; and (c) item bias that occurs when test takers with the same level of the target latent construct obtain different scores on a given item (He et al., 2019; Sireci, 2011).

The increasing diversification of the population among participating countries in ILSAs in terms of jurisdictions, cultures, languages, exposure to curriculum, familiarity with the context or linguistic terms expressed in the test items, previous access to formal schooling, and educational paradigms poses potential threats to the validity of test score's interpretations while increasing the risk of possible unintended consequences such as overgeneralizations of test scores, misuse of test scores, and biased interpretations of the students´ academic skills (Isac et al., 2019; Oliveri et al., 2018; Oliveri & Ercikan, 2011; Oliveri & Lawless, 2018).

In addition, empirical evidence has shown that test scores from multilingual and multicultural versions of a test cannot be assumed to be comparable. Therefore, empirical evidence of construct comparability at both the item and test levels must be collected prior to using data from ILSAs (Drasgow & Probst, 2005; Isac et al., 2019; Oliveri et al., 2018; Oliveri & Ercikan, 2011; Oliveri & Lawless, 2018; Oliveri & von Davier, 2016; Sireci, 2011; Spielberger et al., 2005).

The need of considering the linguistic and cultural background of test takers in the assessment of minorities has been highlighted in several testing guidelines including the AERA et al. (2014) *Standards for Educational and Psychological Testing*, the International Test Commission (ITC) *Guidelines for the large-scale assessment of linguistically and culturally diverse populations* (ITC, 2018), the ITC Guidelines for

Translating and Adapting Tests (second edition) (ITC, 2017), and the *ETS Standards for Quality and Fairness* (Educational Testing Service, 2014). For instance, the guidelines by the ITC address the importance of minimizing the unintended effects of cultural differences that are not relevant for the purpose of the assessment. It is acknowledged that the cultural background of test takers can impact the comprehension of item content as well as the access to test content in general which in turn leads to differences in response processes (Muthén & Asparouhov, 2018; Oliveri et al., 2015; Oliveri & von Davier, 2016). The ITC guidelines point to the importance of ensuring score comparability when assessing populations from linguistically and culturally diverse backgrounds by conducting studies to identify the extent to which test scores are invariant across groups (Byrne et al., 2009; ITC, 2018).

A lack of invariance when assessing students from diverse cultural backgrounds may result in unintended consequences where minority groups can be inaccurately compared against the reference groups and thus, the interpretations of their cognitive abilities and academic achievement based on test scores are biased. In this sense, a major concern when analyzing data from these international assessments is the collection of evidence about the validity of inferences and claims that result after comparing scores across countries. Specifically, the challenge is to guarantee that ILSAs provide fair measures across all the individuals within the target population (Desouky et al., 2013; Oliveri et al., 2015; Oliveri & von Davier, 2016).

1.4 Fairness in Educational Assessments

According to the AERA et al. (2014) *Standards,* test fairness involves four broad aspects: lack of measurement bias, equitable treatment of examinees in the testing

process, access to the constructs measured, and validity of individual test score interpretations for the intended use. Specifically, the lack of bias refers to the extent to which score-based inferences are valid across different groups of test takers and can be achieved by reducing the influence of construct-irrelevant score variance through the evaluation of measurement invariance specially among groups who have been historically discriminated on the basis of their ethnicity, native language or race. Evidence about the consistency of the psychometric features of the assessment instrument across groups and contexts is needed to achieve the intended uses of a measure (AERA et al., 2014; ETS, 2014; Lamm et al., 2019).

In the case of immigrant students, the sources of construct-irrelevant score variance are mostly related to their cultural and linguistic background as well as to the specific life events they experience when trying to adjust to the daily life in the host country such as culture shock, segregation and stress. These experiences can impact their overall perception of life satisfaction across the psychological, social, physical, and cognitive domains. Therefore, the responses from immigrant students will be impacted by their status as immigrants (Byrne et al., 2009; Cheung et al., 2006; OECD, 2017).

1.5 The Importance of Non-cognitive Measures

Given the impact that non-cognitive constructs have on the test performance of immigrant students, most ILSAs include non-cognitive measures that not only serve to optimize achievement estimates and contextualize test results, but also as evaluative tools themselves. Among non-cognitive measures, the measurement of *bullying* has received a lot of attention in the past decades not only due to its high prevalence among general student population, but specifically due to its prevalence among immigrant students. In

fact, empirical evidence has shown that immigrant students are around 33% more likely to be bullied than their native peers in Europe and North America because of their uniqueness in terms of language, culture, ethnicity and physical appearance (OECD, 2017; United Nations Educational, Scientific and Cultural Organization [UNESCO], 2019).

Furthermore, given the role that this non-cognitive construct can play on the initiatives to effectively integrate immigrant students into educational and social systems, its assessment is of great relevance. As noted by the United Nations Educational, Scientific and Cultural Organization (UNESCO), it is not possible to claim that the quality of education is inclusive and equitable for all if there are students who experience violence and bullying in school therefore, school violence and bullying must be addressed to ensure inclusive and equitable quality education while promoting peaceful and inclusive societies (UNESCO, 2019).

Bullying is currently being viewed as an urgent public health problem in several countries around the world mostly due to its serious long-lasting consequences -including suicide, homicide, risk behaviors, crime, academic dropout, development of psychological disorders, and general youth violence- and international collaborative efforts have been made to develop measurement instruments that allow for international comparisons in an effort to better understand the dynamics of bullying and inform initiatives oriented towards its prevention and treatment (Casper et al., 2015; Craig et al., 2009; Marsh et al., 2011; Nansel et al., 2004; OECD, 2017; UNESCO, 2019; Vessey et al., 2014; Wolgast & Donat, 2019).

As with other public health-related problems, early detection and prevention have been proposed as the most efficient way to reduce the prevalence of bullying however, the measurement of bullying represents several challenges including designing a measure that is invariant across subpopulations or cultural groups, providing a definition of bullying, and determining the time frame within which students will be asked to report their experiences. Moreover, a systematic review of published measurement instruments designed to measure bullying by Vessey et al. (2014) showed that most available instruments lacked methodological quality when reporting on the psychometric properties. According to the authors, most of the instruments did not have evidence to support their psychometric soundness and only a few included a report on measurement invariance.

In addition to this problem, evidence has shown that there are several culture-specific factors that can have a differential impact on the overall experience of bullying thus, the nonequivalence of this construct across different cultures and languages is the most common source of error because respondents from different cultural backgrounds tend to interpret the wording of items in different ways and cross-cultural invariance can be particularly difficult to achieve (Desouky et al., 2013; Oliveri et al., 2018; Spielberger et al., 2005; Wolgast & Donat, 2019).

Another non-cognitive construct that has relevance among immigrant students is the *sense of belonging at school*. Its educational relevance lies in that it impacts not only their academic success, but also their psychosocial well-being (Chiu et al., 2016). In fact, empirical evidence has shown that a high sense of belonging at school results in higher psychological health, low rates of delinquency, less chances of dropping out of school

and reduced changes of drug use. In this sense, a higher sense of belonging at school leads to a higher cognitive and psychosocial functioning (Chiu et al., 2012).

Therefore, the sense of belonging at school has been identified as a key factor of overall student success in that it is a psychological state that represents the feeling of being connected to the school. Moreover, the sense of belonging at school experienced by immigrant students is also a key indicator of how well they are being integrated into the school community (OECD, 2015). (Chiu et al., 2012).

However, as with the measures of bullying, the sense of belonging at school can vary depending on the cultural context. For instance, collectivistic countries that emphasize interdependence tend to place high relevance on developing a sense of belonging than cultures that promote autonomy. Therefore, students' cultural background should be considered when evaluating sense of belonging at school (Chiu et al., 2016).

Given the relevance of this non-cognitive construct, PISA includes a measure of sense of belonging at school as part of the student questionnaire. This construct has been related to the quality of teacher-student relations that impacts not only the students' engagement with school, but also their socio-emotional development. In fact, teachers play a key role in promoting a healthy social and emotional development which in turn leads to better academic performance however, teachers can also engage in different types of unfair behaviors with students and evidence has been shown that disadvantaged and immigrant students are more likely to report unfair teacher behavior (OECD, 2017).

In general, more efforts are needed to collect sound empirical evidence to identify the extent to which the inferences to be made from the scores of these instruments are accurate, valid and suitable to be used from different countries throughout the world as

input for the designs and implementation of educational initiatives targeting immigrant students. Particularly, evidence about the extent to which the instruments that measure non-cognitive constructs are invariant across countries and cultures is needed (Casper et al., 2015).

1.6 Measurement Invariance

As previously mentioned, a desired property of a measurement instrument is that the items consistently reflect the target latent variable even when administered to potentially heterogeneous populations in order to make proper inferences about the test takers based on test scores. Moreover, it is expected that the estimated parameters are equivalently applicable across different subgroups in the population. Therefore, measurement invariance is a necessary condition to compare data across different groups, and to make valid inferences from data collected from individuals across several countries (Millsap, 2007). Testing for measurement invariance provides evidence as to what extent the scores from a measurement instrument reflect the underlying target latent variable instead of other cultural or contextual variables that are not relevant for the measurement process which is a necessary requirement to assess the validity of the interpretations made based on test scores (Carter et al., 2014; Desouky et al., 2013; Fischer & Fontaine, 2011; Hox et al., 2012; Oishi, 2006; Oliveri & Lawless, 2018; Rutkowski & Svetina, 2017; Sawatzky et al., 2018; Scherer et al., 2016).

Measurement invariance is therefore based on the notion that the psychometric properties of a measure should be independent of the characteristics of the individual being measured so that the measurement pertains only to the characteristics that are intended to be measured. If measurement invariance is not guaranteed, then comparisons

across cultures or groups cannot be performed because other factors that are irrelevant to the latent trait could be causing the observed differences among groups and these discrepancies between observed and true differences impact the comparability of test scores leading to biased conclusions about the test takers (Desouky et al., 2013; Fischer & Fontaine, 2011; Lee et al., 2011; Millsap, 2007; Oliveri et al., 2015; Schlagel & Sarstedt, 2016).

Given that international comparisons of educational results are only possible if the measurement instruments are invariant across countries, analyses on measurement invariance should be conducted to ensure that the indices and latent variables used for the international comparisons are comparable across the participating countries. However, the general tendency is to assume measurement invariance instead of collecting evidence about the extent to which the measurement instruments are in fact, invariant (Byrne, 2004; Scherer et al., 2016; Wendt et al., 2017).

Measurement invariance is rarely included as a relevant part of the psychometric analysis of international measures (He et al., 2019) and one plausible explanation for this tendency could be related to the lack of statistical techniques that can properly handle both the particular features of data collected through ILSAs, and the challenges that arise when multiple comparisons are to be performed. The statistical analysis of many groups typically involves high levels of measurement non-invariance especially when the groups are culturally diverse, posing several challenges for the implementation of statistical techniques (Muthén & Asparouhov, 2013).

In terms of the features of data from ILSAs, it is well-known that the data are hierarchical and have a nested structure. For instance, a review of studies using PISA data

showed that in general, all the studies highlight the risk of a possible shift in the meaning of the constructs when the levels of analysis (e.g., individual, country) are not considered during the implementation of statistical techniques (Hopfenbeck et al., 2018; Sawatzky et al., 2018).

1.7 Statistical Approaches to the Evaluation of Measurement Invariance

A traditional approach to the assessment of measurement invariance is multiple-group confirmatory factor analysis (CFA) that compares latent variable means, variances, and covariances across groups while holding specific measurement parameters fixed so that they are equal across groups. In this framework, invariance of factor loadings and measurement intercepts (typically referred to as scalar invariance) is required so that factor means can be compared.

However, models with strict invariance are often rejected and modification indexes are then needed to relax some of the invariance restrictions, which need to be estimated manually. Moreover, multiple-group CFA does not consider the nested structure of the data and becomes too cumbersome when applied to a large number of groups as in the case of ILSAs and it is often impractical given the many possible violations of invariance (Asparouhov & Muthén, 2014).

The field of measurement invariance has experienced several changes in the past decades in terms of the development of new methodologies and technical refinements that aim to overcome the well-known limitations of traditional approaches to handle large-scale data from international assessments that typically involve a large number of groups. However, no clear solution has yet been found (van de Vijver et al., 2019).

Recent latent and hierarchical-based approaches to the assessment of measurement invariance, that allow for tests of approximate measurement invariance, have been recently developed (Byrne & van de Vijver, 2017) and among them, the alignment optimization is thought to be a suitable approach to the analysis of measurement invariance because the method (a) estimates models for many groups, (b) automates and simplifies the analyses, and (c) provides a detailed account of parameter invariance for every model parameter within each group (Muthén & Asparouhov, 2014). Moreover, the alignment optimization method can handle, through the incorporation of maximum likelihood estimation, the complex survey features of weights from PISA data that result from using probability proportional to size to sample schools (Muthén & Asparouhov, 2018).

Nevertheless, despite evidence that points to the unsuitability of the traditional approaches for the analysis of measurement invariance to handle data from ILSAs, in the case of PISA -which as noted before is one of the most widely used assessment- most of the documented analyses of measurement invariance rely on traditional approaches, and no research on full measurement invariance has been conducted across cultures (Meng et al., 2018).

In fact, full comparability across countries and subpopulations are not guaranteed and the documentation of PISA 2015 warns users of data about three possible biases in self-reported responses from students: (a) social desirability, (b) reference-group bias that is related to the features of the comparison group, and (c) response-style bias which are thought to operate differently across cultures limiting the comparability of responses across cultures (OECD, 2017). In addition to this issue, a systematic literature review by

Hopfenbeck et al. (2018) showed that most published papers analyzing PISA data have focused on evaluating measurement invariance in the cognitive surveys and only a few focused on self-report questionnaires highlighting the need of more evidence of measurement invariance targeting the non-cognitive constructs measured by PISA.

In general, the criticisms of PISA data have primarily been related to (a) the differential meaning of cultural, social and economic constructs across countries that can make cross-cultural comparisons invalid, (b) evidence suggesting that self-report questionnaires lead to biased inferences due to poor questionnaire design and language ambiguity, (c) high levels of missing data and low reliability in background questionnaires for some countries, and (d) limited interpretation from background questionnaires due to questionable test design (Hopfenbeck et al., 2018).

In conclusion, five major problems are identified in the contexts of ILSAs that are administered to culturally diverse test takers: (a) the tendency to ignore the need of testing for measurement invariance and not including it as part of the psychometric analyses conducted on data from ILSAs, (b) the lack of sound evidence about the extent to which the non-cognitive constructs measured through ILSAs are invariant among immigrant students when compared to their native peers and among immigrant students across countries, (c) lack of evidence about how the measures of exposure to bullying and sense of belonging at school are related and account for the performance of immigrant students in reading literacy, (d) lack of a comprehensive analysis of fairness in ILSAs that target the immigrant student population with respect to specific background variables, and (e) the lack of empirical evidence about the suitability of modern statistical techniques to test for measurement invariance when applied to data from ILSAs.

1.8 General Purpose

To address these issues, the general purpose of this dissertation is to provide evidence about the extent to which the 2018 Programme for International Student Assessment (PISA) provides invariant measures of reading literacy, exposure to bullying, and sense of belonging at school; for immigrant students from diverse cultural and linguistic backgrounds across the countries that host large populations of immigrants.

1.8.1 Specific Purposes

To this end, the specific purposes of this dissertation are to:

1. Determine the extent to which test scores from the reading literacy test are invariant across immigrant and native students within and across countries.
2. Determine the extent to which each of the non-cognitive measures (exposure to bullying and sense of belonging at school) are invariant across immigrant and native students within and across countries.
3. Identify potential sources of construct incomparability.
4. Evaluate the relationship among the non-cognitive measures and the measure of reading literacy to identify the potential role of non-cognitive measures as predictors of achievement in reading among immigrant students. This evaluation will be used to explain the observed performance on reading literacy and to interpret results from the analyses on measurement invariance.

1.9 Research Questions

Given these purposes, the following research questions will be addressed in this dissertation:

1. Are the factor loadings, factorial structures, and item intercepts from the reading literacy test comparable across immigrant and native students?
2. Do the non-cognitive scales measure the same target latent construct across immigrant and native students?
3. What are the potential sources of construct incomparability?
4. Are the test scores from the reading literacy test and the non-cognitive scales comparable across countries?
5. To what extent can the non-cognitive scales predict the performance on the reading literacy test?

1.10 Hypotheses

The hypotheses regarding the research questions are:

1. The factor loadings, factorial structures, and item intercepts from the reading literacy test will not be equivalent across immigrant and native students.
2. The items from non-cognitive scales will not be invariant between native and immigrant students.
3. Construct incomparability among countries will mimic the differences in the cultural dimensions as stated in the cultural model by Hofstede (2011).
4. Construct incomparability within countries will reflect (a) differences in immigration status, (b) restriction of range, and (c) differences in sample size.
5. Test scores from the cognitive and non-cognitive scales will be comparable among countries with similarities in the cultural dimensions stated by Hofstede (2011).

6. The variability in the performance on reading literacy will be explained by the performance on the non-cognitive scales.

1.11 Contributions

This dissertation will contribute to fulfill the existing gaps by providing:

1. Sound empirical evidence about the extent to which test scores from the PISA cognitive and non-cognitive measures are invariant among (a) immigrant students when compared to their native peers, and (b) among immigrant students across countries.
2. Evidence regarding how the non-cognitive measures are related to the test performance of immigrant students in the cognitive domain of reading literacy.
3. A comprehensive analysis of invariance in ILSAs that target the immigrant student population with respect to specific background variables.
4. Sound empirical evidence about the suitability of modern statistical techniques to test for measurement invariance when applied to data from ILSAs.

The next section will provide the theoretical background for this dissertation beginning with a description of the current state of international migration, followed by the implications of that problem for governments throughout the world and how that relates to their educational systems. Then, the educational experiences of immigrant students will be described as well as the role that international large-scale educational assessments play as the primary source of information for the governments around the world.

PISA will be introduced and described in detail along with the challenges that occur when evaluating highly diverse populations and the emphasis will be placed on the difficulties related to the proper establishment of measurement invariance.

Finally, the statistical approaches that have been suggested to address those measurement challenges will be introduced and described. Advantages and disadvantages related to each technique will be provided.

CHAPTER

II LITERATURE REVIEW

2.1 International Migration

In recent years, the number of immigrants around the world has rapidly increased, reaching 258 million by 2017. Around 60% of all international migrants live in Asia or Europe and according to the United Nations, North America hosted the third largest number of international migrants (58 million) followed by Germany and Saudi Arabia (Ratha et al., 2018; United Nations, 2017). In terms of the country of origin, most international migrants are originally from India (17 million) followed by Mexico (11.9 million), Russian Federation (11 million), China (10.1 million), and Bangladesh (7.8 million) (United Nations, 2017) where the internal displacement has increased mainly due to conflict, violent extremism, and natural disasters related to climate change (Ratha et al., 2018).

Traditionally, immigrant populations have been classified into two categories: (a) first-generation immigrants who were born in another country different from the host country, and (b) second-generation immigrants who are native-born individuals with at least one foreign-born parent (Akresh & Akresh, 2011; Duong et al., 2016; Volante et al., 2017). The population rates, as well as the social integration of immigrants into the host country, varies for each of these categories.

In terms of student population, the Global Migration Data Analysis Centre and International Organization for Migration (2018) reported that the international migrant population included 4.8 million international students by 2016. In the case of the United States, immigrant students accounted for 21.5% of the public-school students by 2010

(Bozick et al., 2016) and according to information collected through PISA on 2015, around 23% of the students in the United States have an immigrant background (Organisation for Economic Co-operation and Development, 2016). The population of immigrant students has clearly increased in the past years around the world posing several challenges for the educational systems in the host countries in terms of finding ways to guarantee the effective integration of these students into their educational systems while providing them with resources that are tailored to their particular needs (Bozick et al., 2016).

As migration movements continue to increase throughout the world, governments from the host countries have made international migration a priority in their political agendas given the impact that these movements can have on social cohesion. Governments across the world are now faced with the challenge of developing policies that help maintain a proper social balance where citizens and immigrants can exercise their human rights while being provided with the resources they need to have an active participation in society. The education of immigrant population appears as the first step towards their effective integration into the society of the host country because schools can provide them with the resources they need to have an active role in society of the host country.

2.2 Immigrant Students and the Educational Systems from Host Countries

As previously mentioned, immigrants are one of the fastest-growing demographic populations throughout the world and international governments are faced with the challenge of maintaining social cohesion while developing policies that allow for their proper integration in the society. The education of immigrants is the first step towards

their proper integration since governments that implement effective integration policies can help immigrant students reach their full academic potential -regardless of their linguistic and cultural background- while providing them with further opportunities so that they can have an active participation in the labor market and thus, contribute to the economic growth and development of the host country by paying taxes, contributing to retirement schemes and taking an active role in the local economy (OECD, 2015).

In fact, OECD has suggested that a proper integration of immigrant students into the educational systems is a benchmark of the overall efficacy of social policies within countries and that the effective integration of immigrant students into educational systems is an indicator of both excellence and equity (Rubinstein-Avila, 2016). Accordingly, empirical evidence from studies on immigration policies and schooling administrations has suggested that positive integration policies towards immigrant students are likely to result in educational systems where all students are educated fairly (Arikan et al., 2017). In this sense, the ways in which school systems respond to migration movements can have a large-scale impact on the economic and social well-being of the communities they serve thus, schools are expected to provide immigrant students with the academic resources they need to succeed and become active citizens (OECD, 2015).

Schools are for most school-aged immigrant students, the first social institution they engage with on a regular basis and immigrant families have historically viewed schools from the host countries as agents of social mobility where their children can obtain the resources they need to succeed therefore, schools are one of the most relevant institutions that impact the lives of youth while shaping their overall health and

development. In fact, empirical evidence has shown that the experiences immigrant students have at school can potentially facilitate or hinder their transition into the society of the host country since the features of school systems impact both educational and non-educational outcomes (Crosnoe & Turley, 2011; Dunn et al., 2015; Pivovarova & Powers, 2019).

Recent research focusing on the educational achievement of immigrant students has also shown that an effective integration of this population into the educational systems can result in an increased academic success as students adopt the culture of the host country (Duong et al., 2016). However, that the overall educational experiences of immigrant students are shaped by the interplay of several factors including how their families are received by the host society which is traditionally determined by reactions to race and ethnicity and which at the same time can either promote social inclusion of immigrant students or marginalize them (Crosnoe & Turley, 2011).

The societies of host countries tend to vary according to the relevance they place on collectivism or individualism. Individualism is characterized by attributes of independence, autonomy, self-reliance, uniqueness, achievement orientation and competition whereas collectivism is associated with interdependence with others, desire for social harmony, and conformity with group norms so that behaviors and attitudes are usually determined by norms of the ingroup (e.g., extended family, community) (Green et al., 2005).

Western cultures are typically characterized by individualist traits whereas non-western cultures are mostly characterized by collectivist features as shown by cross-cultural meta-analysis where findings consistently show that people from North America

typically score higher on individualism, personal independence and uniqueness than people from Hong Kong and Japan (Green et al., 2005). Moreover, cross-cultural research has also shown that differences in attitudes, values, behaviors, cognition, communication, socialization, and self-concepts can be described, explained, and predicted through the concepts of individualism and collectivism. Therefore, the socio-cultural context of the host country can have a significant impact on the overall educational experience of immigrant students (Green et al., 2005).

Unfortunately, immigrant students usually experience difficulties in understanding the social and cultural rules that regulate the functioning of the host country but are usually implicit (OECD, 2015). In this sense, the effective integration of immigrant students into the educational systems is challenging -despite the long-term benefits that have been associated to the social inclusion of immigrants- given that it requires school systems to consider their ethnic, cultural, socio-economic, religious and linguistic backgrounds to help them understand and ultimately internalize the culture of the host country (Rubinstein-Avila, 2016). Additionally, efforts towards the training of teachers are needed because typically teachers of immigrant students do not have the training to implement pedagogical approaches to help them achieve their educational goals while meeting the educational standards from the host country (OECD, 2015).

On the other hand, the structure of educational systems also depends on the type of government which in turn, has an impact on the formulation of educational standards that refer to educational outcomes and the determinants of those outcomes. Educational research on this area has shown that the standardization of educational systems can

impact the academic success of immigrant students while non-standardized educational systems tend to marginalize this population (Teltemann & Schunck, 2016).

Empirical evidence has also shown that educational systems -typically from western countries- that implement early tracking of students into academic or vocational programs can increase inequality because students from disadvantaged backgrounds, as it is the case for most immigrant students, tend to be assigned in tracks with lower performance expectations where they are provided with less academically-demanding programs that not only limit their educational development by preventing them from achieving their full academic potential but also create barriers to access higher education and thus, to have high-status professional occupations (OECD, 2015).

In this scenario and as previously mentioned, the structure of educational systems can either facilitate or hinder the inclusion of immigrant students. Therefore, governments are also faced with the task of evaluating their current educational systems to identify how they are impacting the experiences of immigrant students and make the necessary adjustments. Recent evidence has suggested that even though it is usual that immigrant students tend to show a lower academic performance than their native peers, some governments from host countries have managed to reduce this gap. For instance, Germany managed to improve mathematics performance on achievement test among immigrant students by 46 score points in less than a decade. Similarly, first-generation immigrant students in Portugal performed better in 2012 than in 2003 and this improvement was larger than the improvement among native students. Thus, the implementation of integration policies grounded on principles of equality can help immigrant students reach their academic potential (OECD, 2015).

These examples suggest that the traditional observed differences in academic achievement between immigrant students and their native peers could be due to policy-related factors and therefore, might not be reflecting true differences in academic proficiency. In a similar way, Spees et al. (2016) reported that the observed gaps between limited English proficient (LEP) students and their native peers were mainly due to a lack of educational support systems for LEP students. Specifically, schools tend to struggle when it comes to developing effective programs that allow for the successful integration of language and content learning mostly because of limited immigrant-specific resources (e.g., lack of English as second language -ESL- teachers, bilingual staff, ESL courses) that ultimately create language barriers and cultural divisions that hinder academic achievement.

In the case of the United States, given that the fastest growing segments in educational systems over the past decades are those of immigrant students who make up around 25% of the population in the country, immigration is currently an increasingly political issue. For instance, first-generation immigrants have been the center of contemporary political debates about immigration mainly because they are more vulnerable to anti-immigrant initiatives than second-generation immigrant students since their eligibility for citizenship depends on their parent's immigration status. Additionally, immigrant students are not equally distributed across schools and most of them only have the option to attend desegregated schools that in general do not have the resources to meet their specific needs (Duong et al., 2016; Powers & Pivovarova, 2017).

In this scenario, questions about the educational experiences of immigrant students have arose and data from international large-scale assessments (ILSAs) are

being used in an effort to better understand the dynamics of educational systems and their impact on these students to better address their needs and promote their inclusion (Powers & Pivovarova, 2017). A salient advantage of ILSAs is that most of them provide information at different levels. For instance, PISA data includes country data, student achievement data, and school data as well as an index for immigrant background based on three country-specific variables that refer to the students 'country of birth as well as their mother and father´s. These variables are recoded into two categories: (a) country of birth is the same as country of assessment and (b) other, and a general index of immigrant background is calculated from these variables with three categories:

1. Non-immigrant students: students who had at least one parent born in the country.
2. Second-generation immigrant students: students born in the country of assessment but whose parents were born in a different country.
3. First-generation immigrant students: students born outside the country of assessment and whose parents were also born in a different country (OECD, 2017; Rubinstein-Avila, 2016).

By providing this information, governments across countries can better use the data to inform their educational systems and policies regarding immigrant students while obtaining a more comprehensive view of their educational achievement.

In summary, the educational systems from the host countries are shaped by cultural factors that impact how they approach immigrant population and thus, influence the overall educational experiences of immigrant students. Furthermore, empirical evidence has suggested that the typically observed differences in academic achievement

between immigrant students and their peers where the former tend to show lower performance, are likely to be due to the features of some educational systems when the needs of these students are not fully considered and thus, might not be reflecting true differences in academic performance.

Efforts are needed to enhance the educational systems and facilitate the effective integration of immigrant students by identifying their needs and provide them with the proper resources. To this end, ILSAs are a valuable source of information by providing countries with data at different levels so that governments can have a comprehensive view of the educational experiences and outcomes of immigrant students to inform their policies and educational practices, and to ultimately promote their effective integration into the educational systems.

2.2.1 Academic Achievement among Immigrant Students

As previously mentioned, the educational experiences of immigrant students are shaped by several factors linked to the educational systems of the host country (e.g., public educational systems, market-oriented educational systems) and thus, their academic achievement relies on both individual characteristics including their attitudes, socio-economic status, previous academic background; and contextual factors related to the quality and receptiveness of the educational system from the host country (Bozick et al., 2016; Giannelli & Rapallini, 2016; OECD, 2015).

In this sense, and as noted by Pivovarova and Powers (2019), a salient indicator of the adaptation of young immigrants to the society of the host country is their academic achievement. For this reason and given the expanding proportion of second-generation immigrant students, policymakers have become interested in understanding how these students assimilate the language and culture of the host country and how their level of assimilation influences their academic performance (Akresh & Akresh, 2011).

Similarly, educational researchers have been interested over the past decades in analyzing the determinants of education among immigrant students and understanding their achievement patterns. A well-documented finding from educational research targeting immigrant students is the presence of an educational achievement gap between immigrant and native students where the former shows lower academic performance, especially in reading and writing measures (Arikan et al., 2017; Azzolini et al., 2012; Giannelli & Rapallini, 2016; Powers & Pivovarova, 2017; Teltemann & Schunck, 2016). However, recent research has led to mixed findings where some results suggests that immigrant students are more likely to experience conditions that have been typically

associated to low academic performance such as living in neighborhoods with high levels of poverty, attending schools with limited resources, and being socially isolated in class, while other studies show that immigrant students have higher educational achievement than their native peers (Alivernini et al., 2019; Powers & Pivovarova, 2017).

Consequently, educational research has also began to focus on immigrant students who outperform their native peers and are often more successful on achievement tests. This trend has been known as the immigrant paradox because immigrant students enjoy academic advantages over their native peers. For instance, evidence based on the National Education Longitudinal Study (NELS) revealed that students with immigrant parents tend to outperform students with U.S.-born parents on math and science tests, a pattern that is stronger among children from Asian immigrant families. Moreover, these academic advantages are better explained by socioeconomic status so that students from immigrant families who have high socioeconomic resources are the ones who tend to outperform their native peers (Crosnoe & Turley, 2011).

In terms of methodological analyses with respect to immigration, several studies have focused on identifying the factors that are most relevant among immigrant students to predict their academic performance. For instance, a study by Martin et al. (2012) evaluated problem-solving skills, settlement and sociodemographic factors in science and mathematics achievement among immigrant students. The authors performed multilevel hierarchical regression where the dependent variables were science and mathematics achievement and the covariates included problem-solving skills, country, school, socioeconomic status, language background, age of arrival, language at home, and gender. Specifically, the authors formulated a model where problem-solving skills

mediated the relationship between immigrant status and achievement controlling for the sociodemographic variables. Analyses were performed on PISA data collected in 2006 from both first and second-generation immigrant students across 17 countries that had a minimum of 3% of immigrant students in the sample. The authors selected samples that had at least 100 immigrant students and found evidence suggesting that sociodemographic and settlement factors were relevant to immigrant students' achievement.

Another study by Areepattamannil and Kaur (2012) aimed to investigate student and school-level factor associated with science achievement of immigrant and non-immigrant students using PISA data from 2006. The authors implemented two-level hierarchical linear modeling to evaluate the relationships where scientific literacy was included as the dependent variable while 30 student-level variables and 12 school-level variables were included as independent variables. The authors conducted the analyses separately for immigrant and non-immigrant students and found evidence suggesting that student attitudes, engagement and motivation in science and information and communication technology familiarity were significant predictors of science achievement for the two groups of students whereas teacher shortage was associated with science achievement only among immigrant students.

In a similar way, Murat and Frederic conducted a study on 2015 to estimate the impact of potentially influential factors on the sign and magnitude of immigrant gaps. The authors analyzed data from PISA 2006 and 2003, selected 29 countries where immigrants accounted for at least 3% of the student population, and conducted separate analysis for first and second-generation immigrant students. They also conducted mean

imputation method to handle the missing data and implemented linear mixed models to handle the hierarchical nature of the data where regressions were run on the full data set but coefficients on immigrant and native students were kept disaggregated at the country level. The covariates in the model included gender, highest parental occupational status, parents' primary and secondary education, language spoken at home, socioeconomic status, school type, and country of birth; and the outcome variables were science, reading and mathematics. The findings showed that negative gaps were concentrated in the European Union and were affected by school type, student background, country of origin, and language spoken at home. The authors noted that the performance of immigrant students remained substantially below than the performance of native students even after those variables were controlled for.

The diversity among immigrant students is such that some students are at a competitive advantage while others are at a large disadvantage. For instance, immigrant students from East Asia benefit not only from the educational background of their parents but also from the willingness of schools to invest in their education given their historical educational success whereas Latin American immigrant students not only tend to have greater socioeconomic disadvantages, are usually less likely to enroll in postsecondary education, and more likely to drop out of high school; but are also faced with stereotypes by school personnel that could marginalize them in schools (Akresh & Akresh, 2011; Crosnoe & Turley, 2011).

To this regard, van Dijk et al. (2019) found empirical evidence suggesting that teachers' classroom management skills have both a significant direct relation with students' academic motivation for mathematics, and a significant indirect relationship

with students' mathematics achievement so that students with higher levels of motivation have a higher academic achievement. Additionally, empirical evidence has also shown that immigrant students' perceptions of anti-immigrant legislation and discrimination are negatively associated with their overall academic performance (Powers & Pivovarova, 2017).

Another factor that has been found to differentiate the academic outcomes among immigrant students is related to whether they are first or second-generation immigrants so that in general, the latter tend to outperform first-generation immigrants in terms of educational achievement. Moreover, in some cases it has been found that second-generation immigrant students also outperform their native peers (Azzolini et al., 2012). In this scenario, ILSAs like PISA play a crucial role by providing countries with information about the academic performance of these students along with other indicators related to their personal background and individual perceptions that can in turn provide a general view of their educational experiences.

Almost ten years of educational international assessment through the Programme for International Student Assessment (PISA) has consistently shown that in most participating countries immigrant students tend to have a lower test performance when compared to native students. However, some countries including Canada and Australia have successfully integrated immigrant students into the educational systems and as a result, their overall test performance is comparable to that of native students (Cattaneo & Wolter, 2015). In this sense, ILSAs can also provide insights about the possible impact of their current initiatives towards the educational inclusion of immigrant students.

In summary, the academic achievement of immigrant students is impacted by several features of the educational systems from the host countries and by their own individual circumstances and educational experiences. Most students struggle when trying to adopt the culture of the host country and are usually faced with several challenges that are unique to their status as immigrants. Their academic achievement needs to be considered in the light of these factors and ILSAs can provide countries with information to identify specific educational challenges that these students face so that they can better interpret their educational outcomes.

2.2.2 Educational Challenges across Immigrant Populations

As previously mentioned, immigrant students are highly diverse in terms of language proficiency, culture of origin, parental educational level, and previous academic and socioeconomic background; and this diversity has been found to impact their overall academic achievement (Duong et al., 2016). Consequently, immigrant students are faced with more challenges than their native peers in that they must overcome difficulties related to displacement, socio-economic disadvantages, language barriers, the development of a new identity that is consistent to the culture of the host country, and discrimination-related aggressions (Borgonovi, 2018; Chiu et al., 2012).

In terms of language barriers, most immigrant students speak a language different than the language of the host country as it is the case for limited English proficient (LEP) students in the United States who speak a language different than English at home and do not have sufficient mastery of English. In fact, by the year 2000 it was reported that 72% of LEP students spoke Spanish and 44% of them were first-generation immigrants (Spees et al., 2016). Language proficiency is therefore a major issue that if not considered by the

educational systems, could promote school failure among this population of students and prevent them from having the opportunity to reach their full academic potential. Additional educational resources and accommodations should be provided to immigrant students to grant them access to educational and assessments materials so that they can have the same educational experience as their native peers.

Regarding socio-economic disadvantages, most immigrant students come from low-income households (below the federal poverty line) and have previous educational experiences that largely differ from their current educational context. Consequently, these students can only attend schools with poor educational resources where their academic needs are not considered and thus, they are more likely to dropout (Spees et al., 2016). The lack of financial resources has largely been associated to school dropout and early engagement in risk behaviors such as delinquency and substance abuse. Therefore, school systems should also make efforts to identify the financial need among these students and provide them with the basic resources they need to prevent dropout and promote school completion.

Another major challenge for immigrant students is related to discrimination which largely depends on the attitudes that governments hold towards this population. General attitudes towards migration vary across countries, for instance, evidence has suggested that Europe holds more negative views towards immigration while the United States seem to hold more positive views where the majority of the population is in favor of an increase of immigrants (Global Migration Data Analysis Centre & International Organization for Migration, 2018). Furthermore, the contexts of reception of immigrants also vary across racial-ethnic groups so that Latino and African American students are

more likely to face academic failure associated with higher levels of discrimination and negative stereotypes regarding their academic ability than other ethnic groups (Duong et al., 2016).

Additionally, immigrant students are more likely to be exposed to peer aggression than their native peers and thus, are in high risk of undergoing health and adjustment problems including severe anxiety, stress, and depression, and engage in risk behaviors such as aggression, delinquency, and substance abuse among others, which in turn has an ultimately negative impact on their overall academic performance (Duong et al., 2016; Rosen et al., 2013). However, that the sociocultural environment, the sociopolitical indices, and aggregated psychological characteristics influence the occurrence of peer aggression therefore, the educational policies oriented towards the educational inclusion of these students could play a significant role at preventing peer aggressions targeting immigrant students (Nansel et al., 2004; van Hemert et al., 2007).

In summary, and as noted by the OECD (2015), immigrant students are ultimately faced with the challenge of overcoming the effects of trauma that result from the migration experience and this seems to be particularly the case for first-generation immigrants who typically tend to underperform in school and report low levels of life satisfaction (Borgonovi, 2018). Psychological adaptation is, therefore, a key factor when it comes to promote academic potential among immigrant students and it involves a sense of personal and cultural identity that in turn, leads to a sense of overall satisfaction in a new cultural context. For instance, having positive interactions and relationships with the proximate school community (e.g., teachers and peers) has been found to help immigrant students to adjust to their new educational context (OECD, 2015).

Governments and educational policymakers in the host countries committed to promote the educational inclusion of immigrant students should thus consider all the factors that impact their educational outcomes to develop policies that can efficiently facilitate their inclusion. To do so, educational assessments should be used to obtain information about the current state of the students' academic skills as well as information about non-academic variables that are known to have an impact on their academic performance so that they can make informed decisions and provide students with the resources they need to overcome all the challenges they experience.

2.2.3 International Educational Assessments

As previously mentioned, educational assessments play an important role in the educational inclusion of immigrant students by providing governments and educational policymakers with information not only about their academic skills, but about other individual-related variables that impact their educational experiences so that they can make informed decisions about allocation of resources and educational policies targeting this population of students. ILSAs of educational achievement are particularly useful for this purpose and have become more popular in recent years given the increasing number of participating countries.

The most well-known ILSAs include the Programme for International Student Assessment (PISA), the Progress in International Reading Literacy Study (PIRLS), and the Trends in International Mathematics and Science Study (TIMSS). The value of these assessments is associated with the data they provide that serves multiple purposes including monitoring of educational systems, providing policy makers with information about what students know and can do, informing decisions about educational policies,

evaluation, and college admission, and providing stakeholders with an understanding of the context and potential correlates of learning (Akresh & Akresh, 2011; Oliveri & von Davier, 2013; Rutkowski & Rutkowski, 2018; Sandilands et al., 2013). Furthermore, ILSAs provide information about educational achievement over time and across countries and most of these assessments also collect data on background variables that have been related to the observed achievement to promote a better understanding of the results and any observed difference among countries (Wendt et al., 2017).

Scores from ILSAs are thus expected to be compared across countries. However, evidence has shown that test scores cannot be assumed to be comparable across countries and language groups (Oliveri & von Davier, 2013; Oliveri & von Davier, 2016). Therefore, to take advantage of the information provided by these assessments and make valid score-based inferences, evidence must be collected prior to use data from ILSAs about the extent to which scores are comparable (Oliveri & von Davier, 2013).

Achieving score comparability when using data from ILSAs can be particularly challenging mostly due to the high diversity across test takers and countries. For instance, it is very likely to find items functioning differently across examinees with the same level on the target latent trait from different cultural and linguistic groups. The reason for this finding is that test items can be worded or presented in a way that is unfamiliar to linguistic minorities and thus, the difficulty associated with those items is increased beyond the difficulty for test takers from the reference group who have the same level of proficiency. The observed low performance is likely to be misinterpreted as lack of knowledge or abilities instead of a lack of familiarity with the item format which leads to invalid score-based inferences (Oliveri & von Davier, 2013; Oliveri & von Davier, 2016).

In this context, the most common source of differential item functioning (DIF) in data from ILSAs is then the diversity in terms of the examinees' cultural and linguistic backgrounds especially in the current decade where student mobility and the number of immigrants has significantly increased. The contextual factors that impact the examinees' response processes such as language acquisition processes, availability of curricular materials, and the proficiency in the language used in the test should be considered from the initial stages of test development to promote the validity of the inferences that are to be made based on test scores among linguistically and culturally diverse populations (ITC, 2018; Oliveri & von Davier, 2013).

Even though most ILSAs have been carefully designed through systematic evidence-based procedures that aim to promote test validity and fairness, test developers are still faced with some constraints given the increasing diversity in the educational systems among participating countries. For instance, some ILSAs need to be limited to few and specific domains that are common across countries which ultimately limit the scope of the interpretations that can be made from test scores (Oliveri et al., 2018).

Another issue that arises when using ILSAs is related to the proficiency level of participating countries. If the proficiency level of some countries is lower than the targeted proficiency distribution of the test, the validity of test scores' interpretations for the lower performing educational systems can be compromised (Oliveri et al., 2018). In this scenario, the comparability of test scores across countries and subpopulations of examinees cannot be assumed thus, evidence of measurement comparability across language and cultural groups must be collected, and special attention must be given to the

evaluation of the extent to which score-based inferences are valid within and across countries (Oliveri et al., 2012).

For this reason, international guidelines for testing emphasize the need to conduct statistical analyses to determine if a test could lead to biased score-based interpretations when administered to culturally diverse populations and also highlight the need to examine fairness whenever a test is to be administered to diverse populations. Furthermore, test developers are urged to design test items that are free of linguistic and cultural-irrelevant characteristics (Oliveri & von Davier, 2016).

Likewise, according to the ITC (2018), large-scale assessments administered to culturally and linguistically diverse populations should provide specific information about the appropriate and inappropriate uses and interpretations of test scores based on evidence from studies on the invariance of test scores across countries and test takers. Strong assumptions about the invariance of the factor structure across countries are needed to ensure that the underlying factors from the test reflect the same target latent constructs so that educational systems across countries and examinees within countries can then be compared (Davidov et al., 2018; Marsh et al., 2018).

As pointed by Oliveri and Ercikan (2011), score comparability plays a crucial role when analyzing data from ILSAs in that it is a necessary condition to make valid inferences from test scores and to determine the extent to which observed differences in test scores represent true differences in performance across groups of examinees. In a similar way, the Educational Testing Service (2014) suggests that ILSAs should address the needs of nonnative speakers of the test language through comparability studies to reduce potential threats to the validity of score-based interpretations so that the test

takers' knowledge of the target latent construct is disentangled from their linguistic proficiency or other variables that are not relevant for the measurement process (ITC, 2018).

As can be seen, ILSAs are powerful assessment tools that have the potential to impact educational systems throughout the world and they are currently being used by several countries to inform educational policies, curricula, and for decision-making processes in general. The most salient advantages of ILSAs include that they: (a) provide information about the immigrant status of the test takers, (b) provide information about academic performance in several domains, and (c) provide information about non-cognitive variables.

To make use of these assessments, test scores need to be compared both among and within countries. However, given the diversity of the population of test takers, it cannot be assumed that the test scores are comparable thus, statistical analysis must be conducted to determine the extent to which the scores are in fact equivalent and ultimately, to make valid inferences from test scores.

2.2.3.1 Programme for International Student Assessment (PISA). As previously mentioned, PISA is one of the largest ILSAs that also provides information about the educational experiences of immigrant students. It was developed by the Paris-based Organization for Economic Co-operation and Development (OECD) with the aim to provide international comparative educational data suitable to be used for policy-making purposes and in the recent years, it has become a standard metric upon which most educational systems across countries judge their relative performance in terms of internationally agreed targets of quality and equity in education (He et al., 2019; Meng et

al., 2018; OECD, 2016; Rubinstein-Avila, 2016; Volante et al., 2017). The number of participating countries has increased over the years so that data from the 2015 administration featured 35 OECD countries and 37 partner countries for a total of 72 countries throughout the world (OECD, 2017). PISA was initially developed in 1999 with the aim to assess aspects of preparedness for adult life in the sense that information collected through the assessment would provide evidence about the students' abilities for lifelong learning (OECD, 2016). The assessment was designed to be administered to 15-year-old students near the end of their compulsory education and has a literacy approach to identify the extent to which the students are able to apply what they learned in school in real life situations (Hopfenbeck et al., 2018; OECD, 2016; OECD, 2018b; Pivovarova & Powers, 2019). In terms of test and item design, PISA is a computer-based assessment mainly delivered in computer format that includes three types of item formats: multiple-choice, short answer and extended response. The items are matrix-sampled across booklets therefore, each student is presented with only a subset of items and then measures of group performance are obtained by aggregating data across subsamples and item subsets (OECD, 2017; Oliveri & von Davier, 2011).

Regarding the sampling design, PISA implements a two-stage stratified sampling design to select schools and students randomly so that in the first stage of the sampling, a number of schools are randomly samples in each country using probability-proportional-to-size sampling and then, in the second stage, eligible students are randomly sampled from each selected school. Sampled students receive a final weight that incorporates the school weight which corresponds to the inverse of the probability that the school is selected, and the within-school student weight that corresponds to the inverse of the

students' probability of selection (Cattaneo & Wolter, 2015; OECD, 2018b; Oliveri & von Davier, 2011).

Three general core school subjects are assessed in PISA: science literacy, reading literacy, and mathematical literacy (OECD, 2016). PISA is administered on a three-year cycle and in each cycle, students are assessed on mathematics, science and reading but, among these three areas one is selected as the major domain for a particular year meaning that more emphasis is placed on the assessment of that area while the other two areas are evaluated as minor domains meaning that they are assessed less thoroughly (Oliveri & von Davier, 2011).

In addition to the three core academic domains, PISA also administers a background questionnaire to students to collect information about family background, economic, social and cultural capital, students' attitudes towards learning, students' perceptions of the school environment, habits, life outside school, and motivation, among other personal variables (OECD, 2016; Oliveri & von Davier, 2011; Powers & Pivovarova, 2017).

Specifically, the student questionnaire is designed to be completed within 35 minutes and it assesses:

- Students and family backgrounds.
- Attitudes towards learning, habits, and life in and outside school, family environment.
- Quality of school's human and material resources, public and private management and funding, decision-making processes, staffing practices, school's curricular emphasis and extracurricular activities offered.

- Context of instruction: institutional structures and types, class size, classroom and school climate and science activities in class.
- Learning: students' interest, motivation, and engagement (OECD, 2017).

Given that educational policies throughout the world are increasingly considering students' well-being as a key factor within the educational systems that needs to be addressed to promote overall student success, PISA recently included a multidimensional measure of wellbeing as part of the student questionnaire that consists of the following dimensions:

1. Psychological: includes students' sense of purpose in life, self-awareness, affective states, and emotional strength. This dimension is measured through students' reports on motivation for achievement and schoolwork-related anxiety.
2. Social: related to the quality of social life in terms of relationships with family, peers, and teachers. This domain is measure through self-reports on sense of belonging at school, exposure to bullying and perceptions of teachers' fairness.
3. Cognitive: includes the cognitive foundations students need to have an active participation in society like the proficiency to use academic knowledge to solve problems and critical thinking. This domain is measured through the competency domains of science, mathematics, reading, collaborative problem solving, and financial literacy.
4. Physical: related to health and adoption of healthy lifestyles and it is measured through engagement in physical activity and eating habits (OECD, 2017).

By including information on these dimensions, PISA provides international governments with relevant indicators to (a) better interpret the students' educational achievement, (b) inform educational policies, and (c) identify ways to promote students' wellbeing.

Additionally, information about equity in education can be obtained by: (a) examining the variation in the distribution of student outcomes as a way to assess the inclusiveness of school systems, (b) evaluating the impact of students' backgrounds on their school outcomes as a way to assess fairness, and (c) exploring how access to educational resources and the incidence of sorting practices vary across students from different backgrounds to identify factors that could mediate their association with performance (OECD, 2016; Volante et al., 2017).

However, in the recent years results from PISA have been frequently used to study achievement gaps among cultural groups and across countries given the disparities that have been reported using PISA data from past years. For instance, analyses from the 2006 PISA data showed that the differences in test performance between immigrant and native students was particularly pronounced in Austria, Belgium, Denmark, France, Germany, the Netherlands and Switzerland while the smallest differences were found especially in Australia, Canada and New Zealand (Cattaneo & Wolter, 2015; OECD, 2016; Volante et al., 2017).

These findings have stimulated research focused on the populations of immigrant students throughout the world. Governments and educational policy makers are interested in comparing both cognitive and non-cognitive measures from PISA between immigrant and native students and comparing results across countries to shed light on their current

educational policies and find ways to promote the effective inclusion of immigrant students into their educational systems. But as previously mentioned, comparing data from ILSAs involves several challenges given the diversity among test takers and countries for example, PISA participating countries are highly diverse in terms of their immigration policies and cultural and linguistic backgrounds, which compromises the comparability of test scores (Cattaneo & Wolter, 2015).

Currently, the procedures used to establish score comparability are based on item response theory (IRT) and most ILSAs provide "senate weights" (where the weights from each country sum up to a constant usually 500) to ensure participating countries contribute equally to the estimation of item parameters; by doing so, it is assumed the international samples along with the estimated international parameters are a proper representation of the functioning of items in each participating country (Oliveri & von Davier, 2011). According to Oliveri and von Davier (2011) it is not possible to assume that the test items will fit well in all countries given the cultural diversity and other country specific factors in fact, evidence has shown that using international parameters for all test items does not lead to accurate representations of the parameters for each country. Consequently, a major concern among researchers analyzing PISA data is the collection of evidence about the extent to which the measures are invariant across immigrant and native students given that the validity of the comparisons will depend on the extent to which the data are equivalent across groups (Hopfenbeck et al., 2018).

PISA has recently made efforts to address the issue of measurement invariance by using a two-parameter item response theory model to examine the comparability of the items and allowing country-specific item parameters when the parameters show poor

item fit. However, there are still many concerns as to the extent to which those measures provide appropriate estimates of population statistics (Braun & von Davier, 2017; He et al., 2019).

In summary, PISA is one of the most widely used ILSAs by governments throughout the world and it is currently being used as a standard against which countries can evaluate their educational systems. The assessment is, therefore, a powerful tool for decision-making processes that impacts educational policies by providing information about academic achievement and about other non-cognitive variables that are related to educational performance. Furthermore, PISA provides information about immigrant students that can shed light about fairness across educational systems.

In this sense, results are expected to be compared across subpopulations of students and across countries however, given the high diversity among students and countries in terms of their cultural and linguistic backgrounds, test scores cannot be assumed to be equivalent and thus evidence about the extent to which scores are invariant across examinees and countries should be collected to guarantee the validity of the inferences to be made from test scores. Despite the efforts that PISA has made to address the problem of measurement invariance, empirical evidence still suggests that the measures are not fully invariant.

2.2.3.1.1 Cognitive-related Constructs. PISA evaluates three major school subjects: (a) science literacy, defined as the ability to engage with science related issues and ideas of science as a reflective citizen, (b) reading literacy, defined as the ability to understand, use, reflect on, and engage with written texts to achieve goals, develop knowledge and participate in society, and (c) mathematical literacy, defined as the ability

to formulate, use and interpret mathematics in a variety of contexts as well as the ability to use mathematical concepts, procedures, facts and tools to describe, explain and predict phenomena (OECD, 2016).

Traditionally, educational researchers have been interested on analyzing these cognitive constructs and recently, several studies have focused on the test performance among immigrant students. The interest on this population started to increase as the empirical evidence began to suggest an achievement gap between immigrant students and their native peers. For example, in the case of limited English proficient (LEP) students, empirical evidence has traditionally shown that they achieve poorer academic outcomes when compared to their native peers. Moreover, national trends suggest a linguistic achievement gap where approximately 71% of LEP students obtain lower scores on standardized math and reading tests than their native peers, and they are also less likely to complete high school and enroll in college (Spees et al., 2016). Additionally, recent evidence has suggested that the achievement gaps seem to be more prominent in some countries. For instance, according to Borgonovi (2018), low academic performance among immigrant students is particularly pronounced in Austria, Belgium, Denmark, Finland, Germany, Iceland, Luxembourg, Slovenia, Sweden, and Switzerland.

However, these findings are not consistent and some studies on the academic performance of immigrant students have shown mixed results that provide evidence about the so-called immigrant paradox according to which some foreign-born students of Latino, East Asian, Filipino, and European descent obtain higher grades in mathematics and English than their native peers. In an attempt to explain this paradox, it has been suggested that immigrant students who successfully integrate to their school environment

while maintaining the values and beliefs from their culture of origin, can show an academic advantage (Duong et al., 2016).

These findings highlight the significant role that the cultural background of test takers can have on their test performance. As noted by Sandilands et al. (2013), the cognitive processes are likely to be affected by language and culture in fact, culturally distinct groups have particular patterns of thinking and learning that when not controlled, become sources of DIF. Thus, it is important to identify the extent to which the cognitive processes that are being measured on a test might be different for distinct language and cultural groups of examinees because failing to do so could have serious implications in the validity of the inferences to be made from test scores. Moreover, these findings suggest that the differences in test performance do not necessarily reflect true differences in the underlying latent trait being measured but instead could be showing differences that are related to the status as immigrant (Gorges et al., 2017).

In a similar way, Ghorbandordinejad and Bayat (2014) have pointed that immigrant students have difficulties understanding the meaning of texts when they are not familiar with the culture from the host country which could in turn, impact their performance on reading comprehension tasks.

In conclusion, the cognitive-related measures have traditionally been the focus of research on ILSAs and most findings suggested a performance gap between immigrant students and their native peers. However, recent research has shown mixed results where in some cases immigrant students outperform their peers which in turn highlights the role that the cultural and linguistic background of the test takers has on their performance. In the light of this evidence, the performance on cognitive-related constructs should be

analyzed along with non-cognitive measures to better interpret the nature of the observed differences in test performance.

2.2.3.1.2 Non-Cognitive Measures. As previously mentioned, an added value of ILSAs is that not only do they provide measures of academic achievement but measures of background information such as affective and behavioral measures, teacher beliefs and practices, and the principals' perspectives on school safety, among others. The relevance of this background information is that it can be used to contextualize and better understand the overall academic achievement especially in the case of subpopulations of immigrant students (Rutkowski & Rutkowski, 2018). However, the traditional focus when analyzing data from ILSAs has been on the achievement measures and only until recently the interest has begun to expand towards the measurement of non-cognitive measures (He et al., 2019; Rutkowski & Svetina, 2017).

ILSAs are increasingly including non-achievement measures that are assumed to have an impact on academic achievement and thus, could contribute to the understanding and interpretation of the academic outcomes increasing the likelihood that the inferences and interpretations made from test scores are valid. Moreover, data on student-related variables is being used to inform educational practices, promote efficient learning strategies, and improve the overall educational systems (OECD, 2015; OECD, 2017; Rutkowski & Svetina, 2017).

As previously mentioned, the educational experiences of the students depend on the features of the cultural context of the countries where they live. For instance, eastern and western cultures are known for their complexities and marked differences especially in terms of collectivism and individualism since individualist cultures focus on the

pursuit of personal goals, autonomy, and independence from others, whereas collectivist cultures focus on the preservation of relationships, social harmony, and independence (Meng et al., 2018). Typically, these cultural features determine the individual experiences of students as well as the way in which they respond to test items specially those intended to measure non-cognitive constructs (Meng et al., 2018).

Understanding the ways in which the cultural context impacts the response to test items can be particularly complicated in the case of immigrant students given that their experiences are shaped by both the culture from their country of origin and the culture from the host country. However, the analysis of their performance on the non-cognitive measures can help identify not only their educational experiences but their adoption of the culture from the host country which in turn can provide an idea about their inclusion into the educational systems from the host country.

Additionally, governments throughout the world are increasingly making efforts to understand how their educational policies impact immigrant students. For example, in the United States, school accountability pressures under the No Child Left Behind (NCLB) Act of 2001 and Every Student Succeeds Act (ESSA) of 2015 increased as the number of immigrant families significantly increased across the country and as a result, educational policy makers are dedicating efforts to understand how the demographic shifts in terms of immigrant population can impact the wellbeing of Limited English Proficient (LEP) students (Spees et al., 2016).

In general, empirical evidence has shown that individual and family policies along with sociocultural and demographic variables impact the academic achievement of immigrant student. Moreover, it has been reported that the factors associated with

educational attainment among immigrant students include psychological-related variables, the parents' educational background, language skills, social interactions with the school community, and the composition of schools (Borgonovi, 2018; Volante et al., 2017). Failure to obtain accurate measures of these variables could have several implications for immigrant students such as academic failure, increased likelihood of dropping out of school, increased prevalence of psychological distress, increased tendency to engage in high-risk behaviors; and also for educational institutions in terms of inadequate allocation of resources, decreased retention rates, failure to promote educational inclusion, and lack of accurate information for decision-making processes, among others.

In this scenario, the relevance of non-cognitive variables is evident which is why the most widely used ILSAs include a section to measure these variables. In the case of PISA, its most recent versions include indicators of students' well-being which is defined as "the psychological, cognitive, social and physical functioning and capabilities that students need to live a happy and fulfilling life" (OECD, 2017, p. 35).

The assessment of students' well-being in PISA 2015 includes sections on (a) performance at school and life satisfaction, schoolwork-related anxiety which is assumed to be negatively associated to performance in science, mathematics and reading, (b) students' motivation to achieve, (c) expectation of further education, (d) students' social life at school, (e) bullying, (f) parents and the home environment, (g) parents' interest in their child's school life, (h) physical exercise and eating habits, (i) working for pay or in the household, and (j) use of information and communication technologies (OECD, 2017).

In terms of the item format for these non-cognitive measures, PISA 2015 implements a Likert-type scale for items in the student questionnaire so that students indicate their level of agreement with each statement (He et al., 2019). Among the non-cognitive measures, anxiety and sense of belonging at school have been found to be related to the performance on cognitive measures and recently, the measures of bullying have been the focus of attention given its high prevalence across schools throughout the world. Given that these variables can play a crucial role in the educational experiences of immigrant students, they will be described next in more detail.

2.2.3.1.2.1 Bullying. Bullying has become a serious global public concern across schools given its high prevalence and the significant long-term adverse consequences that it brings for both the victims and the bullies. This phenomenon not only has gained the attention from governments all over the world but has also prompted a significant amount of research over the past years (Hussein, 2010; OECD, 2017; Volk et al., 2017). According to the UNESCO (2019), currently almost one in three students has been bullied by their peers at school at least once in the last month.

Bullying can be defined as an intentional action aiming to inflict physical and psychological harm on another person that takes place on a complex interplay of dominance and social status and involves repetition, intention to harm, and unequal power between the bully and the victim. Moreover, bullying involves several aggressive social behaviors including name-calling, extortion, physical violence, group exclusion, and damage to property among others (Alivernini et al., 2019; Casper et al., 2015; Craig et al., 2009; Hussein, 2010; Marsh et al., 2011; Nansel et al., 2004; OECD, 2017; UNESCO, 2019; Volk et al., 2017; Wolgast & Donat, 2019).

A key concept in the context of bullying is victimization which has been described as a systematic exposure to peer maltreatment. Victimization can be overt when the target individual experiences physical and/or verbal attacks, or social when the target individuals experience attacks directed to harm their social status, relationships, or self-esteem such as social exclusion (Rosen et al., 2013).

Adolescents can take part in bullying performing different roles such as perpetrator, perpetrator's assistant, reinforcer, bystander, victim's defender, and victim. However, the most adverse consequences from bullying impact perpetrators and victims so that in the case of perpetrators the risk for maladaptive development into adulthood is higher, and they are also more likely to engage in risk behaviors such as smoking, drinking alcohol, substance abuse, engagement in criminal behavior and tendency to suffer from internalizing problems like high depression, low self-esteem, low mood level and physical complaints. Moreover, perpetrators are more likely to have suicide ideation and make suicide attempts. The victims on the other hand, tend to suffer from anxiety, are more likely to isolate from social interactions, show low academic performance, and can also have high levels of depression and engage in risk behaviors like smoking and alcohol consumption while being in high risk of suicide (Flouri & Papachristou, 2019; Wolgast & Donat, 2019).

Bullying can also be conceptualized through a social-ecological diathesis-stress model to explain how the biological and psychological predispositions along with the environmental stressors contribute to the development of mental and physical disorders. In this sense, individuals who are exposed to stressors like bullying can develop disorders

which can vary on intensity depending on their personal resources and their access to contextual resources (Wolgast & Donat, 2019).

The most common adverse consequences of bullying that can last throughout adulthood include school dropout, poor academic performance, increased risk for depression and anxiety, low self-esteem, social isolation, alteration of eating habits, use of illegal substances, and suicide. Students who have specific features in terms of age, physical appearance, gender, and ethnicity are more likely to become either a bully or the victim of a bully (OECD, 2017; UNESCO, 2019; Vessey et al., 2014). For instance, students who are perceived as "different" from the general school population are more likely to be bullied and to feel like an outsider. In fact, international surveys have shown that physical appearance is the most common reason for being bullied while race, nationality and/or skin color are the second most common reason. Moreover, immigrant children from low-income families are more vulnerable to bullying specially in host countries where the attitudes towards immigration are negative (Alivernini et al., 2019; UNESCO, 2019).

Empirical evidence has shown that victims of bullying have poor emotional adjustment, low quality relationships with peers. Bullies on the other hand, tend to show poor school adjustment and frequent alcohol consumption therefore, the consequences of bullying reach not only the victims but the perpetrators as well. The odds of weapon carrying are higher for students involved in bullying either as bullies or victims than for the rest of the student population (Nansel et al., 2004).

Different types of bullying have been identified including: (a) physical bullying, characterized by repeated physical aggression; (b) psychological bullying, that includes

verbal abuse, emotional abuse and the manipulation of social relationships to harm or exclude the person being victimized; (c) sexual bullying that refers to being made fun of with sexual jokes or comments; and (d) cyberbullying, that refers to being bullied by messages or pictures and also refers to being treated in hurtful ways through mobile phones or online. Moreover, these types can overlap with one another (Casper et al., 2015; Craig et al., 2009; OECD, 2017; UNESCO, 2019).

Given the increasing incidents of bullying across educational institutions, awareness on this topic has resulted in national mandates to act and intervene on bullying-related incidents and several efforts have been made to approach the problem through preventive initiatives and interventions. However, prior to the implementation of these initiatives it is important to evaluate the suitability of the measurement instruments that are being used to collect data on bullying and that are expected to inform the design and development of intervention strategies. In this sense, empirical and sound evidence about the psychometric properties of these measurement instruments should be collected to identify the extent to which the inferences to be made from the scores are accurate and valid (Casper et al., 2015; Rosen et al., 2013).

However, that the evaluation of bullying poses significant measurement challenges. For instance, given that some reliability indexes as Cronbach's alpha have highly restrictive assumptions implying that every item measures the target construct to the same degree and contains the same amount of item-specific variance, the use of those indexes for bullying scales targeting diverse subgroups could lead to inaccurate estimates of the internal consistency which in turn could lead to biased inferences from the test scores (Casper et al., 2015).

Additionally, items within bullying scales can vary in terms of their psychometric functioning across developmental stages and contexts (i.e., bullying within a classroom versus bullying on a playground) and also in terms of the experiences and expressions of anger therefore, evaluations of measurement equivalence are highly encouraged (Casper et al., 2015; Spielberger et al., 2005). To this regard, cross-cultural research has shown that the issue related to the psycholinguistic equivalence of terms used to explain bullying-related behavior remains within and between countries. Moreover, evidence has shown that the interpretation of the types of bullying also varies across countries and in general, instruments to measure bullying are sensitive to socio-cultural differences including cultural norms, socioeconomic inequality, and cultural values such as individualism and collectivism so that individualist societies report less overall victimization but larger indexes of relational bullying than collectivist societies (Samara et al., 2019).

On the other hand, empirical evidence has pointed that international measures of bullying have some limitations such as a focus on the extreme ends of the continuum of bullying behaviors, a lack of efforts to reduce social desirability, and content underrepresentation (Marsh et al., 2011). The OECD has made efforts to address these challenges by providing a clear definition of the construct domain in PISA and evaluating the extent to which the measure is invariant.

The bullying measure provided by PISA is included as a section within the students' well-being questionnaire where bullying is defined as a systematic abuse of power that can be physical, verbal or relational (characterized by social exclusion and diverse forms of public humiliation and shaming) (OECD, 2017). The section includes

six questions to measure bullying through self-reports from the victim's perspective and provides an index of exposure to bullying summarizing the students' answers to the questions. The index is standardized with a mean of zero and standard deviation of 1 across countries so that positive values correspond to students who reported to be bullied more frequently than the average student while negative values correspond to students who were less exposed to bullying (OECD, 2017). Specifically, students are asked to report on the frequency with which they have experienced each of the experiences described in each of the six statements during the past 12 months and they are provided with the following response options: never or almost never, a few times a year, and a few times a month or once a week or more (OECD, 2017).

Finally, regarding the measurement invariance of the scale, three levels of invariance were analyzed: (a) configural invariance, to determine if the same construct was being measured with the same indicators for two or more populations; (b) metric or equal slopes invariance, to test if the factor loadings are statistically equivalent in addition to configural invariance; and (c) scalar or equal slopes and thresholds invariance, to test if all the thresholds are statistically equivalent in addition to metric invariance. The documentation of PISA 2015 reports that partial invariance was achieved for the scale where at least three of the items were fixed across all countries and three were allowed to vary. Model fit was measured through the comparative fit index (CFI) and the root mean square error of approximation (RMSEA). However, given that full invariance was not achieved, the documentation states that caution must be taken when interpreting cross-country analysis based on the bullying scale (OECD, 2017).

2.2.3.1.2.2 Sense of Belonging at School. Sense of belonging at school has been identified as a key factor of student success in that it is a psychological state that represents the feeling of being connected to the school and it has also been associated to cognitive and psychosocial functioning (Chiu et al., 2012). Sense of belonging at school can be defined as a psychological state where students view their schools as essential to their overall well-being and it can be manifested in an active engagement in both academic and non-academic pursuits as well as in the relationships that students establish with school staff and their peers (Chiu et al., 2016).

Empirical evidence has shown that a high sense of belonging at school results in higher psychological health and positive affective states, low rates of delinquency, less chance of dropping out of school, and reduced changes of drug use. In this sense, a high sense of belonging at school leads to a higher cognitive and psychosocial functioning thus, the educational relevance of students' sense of belonging at school lies in that not only does it impact their academic achievement but their psychosocial well-being as well (Chiu et al., 2012; Chiu et al., 2016; OECD, 2015).

As it is the case for most non-cognitive constructs, the sense of belonging also varies depending on cultural context. For instance, collectivistic countries that emphasize interdependence tend to place high relevance on developing a sense of belonging than cultures that promote autonomy. Consequently, students from collectivistic cultures show high sensitivity to their peers' behaviors, recognize and adopt well-discipline classmates' model behaviors, receive more positive feedback from the teachers, feel more successful at school and thus, have a high sense of belonging at school. Therefore, students' cultural

background should be considered when evaluating sense of belonging at school (Chiu et al., 2016).

In the case of immigrant students, this construct is very likely to be affected since they have a cultural and linguistic background different from that of the host country therefore, their sense of belonging will depend on how much they have adopted the culture from the host country and how much they have been effectively integrated into the educational systems by their immediate school community. In this scenario, the sense of belonging at school experienced by immigrant students is a key indicator of how well they are being integrated into the school community and thus, it can provide international governments with a criterion to judge the effectiveness of their efforts towards the educational inclusion of these students (OECD, 2015).

Moreover, among the school community, teachers can particularly influence the students´ sense of belonging at school and their academic engagement through their interactions and the development of a supportive classroom climate. However, most immigrant students face economic hardship and therefore, have to attend poorer schools where they are likely to find less skilled teachers who usually fail to provide them with the support and resources they need to achieve academic success (Chiu et al., 2012).

Unfortunately, immigrant students throughout the world are likely to experience alienation from the educational systems and thus, they often feel they do not belong at school. Moreover, according to empirical evidence, they are also more likely to report unfair teacher behavior (OECD, 2015; OECD, 2017). The assessment of this construct is highly relevant given its association with educational policies, academic achievement, and overall psychosocial wellbeing. PISA provides a measure of students' sense of

belonging at school within the questionnaire of students' well-being that evaluates feelings of social acceptance and attachment to the school community (He et al., 2019; OECD, 2017).

Analysis of PISA data from 2015 have shown that 23% of the students in OECD countries are immigrants and on average these students showed lower performance while reporting a weaker sense of belonging at school, less satisfaction with life and higher level of anxiety when compared to nonimmigrant students (Borgonovi, 2018). These findings confirm that immigrant students seem to struggle to develop a sense of belonging at school however, and despite the relevance of this construct in terms of academic success and psychosocial wellbeing, little research is available on this topic (Chiu et al., 2016).

More efforts are needed not only to further explore the impact of this construct on the overall educational experiences of immigrant students but also to collect sound empirical evidence about the quality of the available assessment instruments that measure this construct and the extent to which they provide measures that are invariant across highly diverse populations.

2.3 Measurement Invariance

Educational researchers usually implement several strategies such as back translation and cognitive interviews -among others- to ensure that the measurement instruments they use to collect information can be transferred across cultures. However, it is likely that these instruments can lead to biased interpretations of test scores because (a) cultural systems determine the meaning and characteristics of cognitive processes and psychological constructs, (b) methodological biases such as translation biases, perception

of response styles, lack of familiarity with testing procedures, and construct underrepresentation can affect specific items or the whole instrument violating the requirements for measurement equivalence, and (c) the lack of generalizability of the constructs from the individual level to the national or cultural level can impact the observed differences on test scores (Samara et al., 2019).

In this sense, the relationships between the items and the underlying latent construct are likely to change across respondents from diverse cultural groups therefore, the extent to which the measures are invariant across subgroups of respondents must be evaluated (Samara et al., 2019).

As previously mentioned, measurement invariance is a necessary requirement to (a) compare data from ILSAs within and among countries, (b) identify meaningful cultural differences, and (c) ensure that the items measure the underlying target latent construct in the same way across culturally diverse populations (Casper et al., 2015; Fischer & Karl, 2019; Halamová et al., 2019; Hussein, 2010; Marsh et al., 2018; Martin et al., 2019; Rutkowski & Rutkowski, 2018; Rutkowski & Svetina, 2014; Sandilands et al., 2013; Seddig & Lomazzi, 2019). Thus, studies on measurement invariance provide evidence about the quality of measurement instruments to improve the validity of the interpretations to be made from test scores (Cieciuch et al., 2014).

Measurement invariance is a property of a measurement instrument that is achieved when the instruments measure the same target latent construct in the same way with the same degree of uncertainty across groups of respondents regardless of group membership (Davidov et al., 2018; Isac et al., 2019; Martin et al., 2019; Oliveri & von

Davier, 2013; Rikoon & Midkiff, 2018; Schlagel & Sarstedt, 2016; Verdín & Godwin, 2017). Measurement invariance (MI) can be defined in terms of a conditional probability

$$\mathrm{P}(\mathrm{X}|\mathrm{W},\mathrm{V}) = \mathrm{P}(\mathrm{X}|\mathrm{W}) \qquad (1)$$

where X is the $q \times 1$ vector of random variables representing scores on observed measures. W is an $r \times 1$ vector representing the target latent variables for X.

V represents a $s \times 1$ vector of measured variables defining person characteristics of interest that should be irrelevant to X once W is considered. In some studies, $\mathrm{s} = 1$ and V is a scalar group identifier that defines demographic variables like ethnicity. Thus, measurement invariance of X with respect to W and V holds if the equality in (1) holds.

For all $\mathrm{X}, \mathrm{W}, \mathrm{V}$ where $\mathrm{P}(\mathrm{X}|\mathrm{W})$ is the conditional probability function for X given W. This probability can be expressed either as a discrete conditional probability when X is discrete or as a conditional probability density function if X is continuous (Millsap, 2007).

Research on measurement invariance is typically conducted to determine if the individual-level factor structure of the target construct being measured and its variance are the same across groups of test takers (Muthén et al., 1997; Rosen et al., 2013). Failure to establish MI particularly when analyzing data from ILSAs can lead to biased estimates of the target latent construct where observed differences in the target latent trait cannot either be isolated from differences due to group membership or attributed to group differences in the latent construct as they might be the result of the way the measurement instrument operates across groups (Byrne & van de Vijver, 2010; Davidov et al., 2018; Halamová et al., 2019; Hussein, 2010; Jak et al., 2014; Kline, 2016; Marsh et al., 2018; Martin et al., 2019; Williams et al., 2018). Therefore, the assessment of bias is a

requirement in cross-cultural and international research to guarantee that the comparative inferences across cultures are valid (Sireci et al., 2005).

The term bias refers to systematic differences in the outcomes from a measurement instrument that are not the result of differences in the target latent construct but are due to other cultural-related variables that are irrelevant to the construct being measured (i.e., translation issues, item irrelevance). Bias can be classified into three categories:

1. Construct bias. The main cause of this type of bias is differential appropriateness of construct-relevant behaviors across cultures and it leads to construct nonequivalence. Sources of this type of bias include inadequate sampling of construct-relevant behaviors and inadequate coverage of the aspects related to the target construct.
2. Method bias. This category includes bias due to factors related to the methodology of the research study including sample selection, test administration (e.g., lack of standardized procedures, ambiguous instructions, differential familiarity with administration material, test administrator effects) and characteristic of the measurement instruments.
3. Item bias. Also known as differential item functioning (DIF) that occurs when test takers with the same level of the target latent construct being measured have different scores/responses on a given item or set of items due to cultural differences. The most common causes of DIF include poor translations and low familiarity of item content (van de Vijver & Leung, 2011; van de Vijver & Poortinga, 2005).

The most common factors that can affect the comparability of measures and increase bias include the test takers' cultural background, the level of proficiency in the language of testing, translation issues, social desirability, curriculum coverage, and familiarity with test content and format. Moreover, these factors not only affect score comparability across countries but across subpopulations within countries (Lomazzi, 2018; Oliveri & von Davier, 2011).

Construct comparability is likely to be compromised in ILSAs given the cultural diversity among test takers who typically have different degrees of familiarity, knowledge and experience with the cultural beliefs, values and practices of the cultural group for which the initial assessment was developed (Oliveri & Ercikan, 2011). Construct comparability can be compromised in different degrees therefore, three levels of MI have traditionally been distinguished and each level is defined by the parameters that are constrained to be equal across groups: (a) configural invariance where it is required that the same latent constructs are measured by the same items across groups and it is achieved if the latent construct has the same factorial structure across groups, (b) metric invariance where factor loadings are set to be equal across groups, and (c) scalar invariance that allows for the comparison of covariances and unstandardized regression coefficients across groups and it is achieved when the intercepts of the indicators are the same across groups in addition to the equality of item factor loadings (Byrne et al., 2009; Byrne & van de Vijver, 2010; Cheung et al., 2006; Cieciuch et al., 2014; Davidov et al., 2018; Fischer & Fontaine, 2011; Kline, 2016; Lee et al., 2011; Lomazzi, 2018; Oliveri & von Davier, 2011; Scherer et al., 2016; Seddig & Lomazzi, 2019; van De Vijver & Poortinga, 2005; Williams et al., 2018). Factor means from ILSAs can only be validly

compared across countries and subgroups of test takers if scalar invariance holds (Seddig & Lomazzi, 2019; van de Vijver & Leung, 2011; van de Vijver & Poortinga, 2005).

In summary, educational researchers are faced with methodological challenges when analyzing data from ILSAs, including how the cultural differences across countries might impact statistical modeling and how to compare countries that could differ significantly from one another (Rutkowski & Rutkowski, 2018). Therefore, an evaluation of MI should be conducted before comparing data from ILSAs to determine the extent to which observed differences in performance across countries and examinees are due to construct incomparability. Also, given that score comparability can be threatened at the item and test levels; it is highly recommended to conduct it at both levels (Oliveri & Ercikan, 2011; Oliveri & von Davier, 2013).

Traditional statistical analyses to test for measurement invariance involve a set of nested tests implemented from the least to the most restrictive. In this sense, the first analysis tests for configural invariance where the aim is to determine if the number of latent variables, the pattern of factor loadings and the measurement errors that underlie the set of test items are the same across countries or groups of examinees. The second test in the hierarchy is the test for metric invariance that provides evidence about the extent to which the pattern and value of the factor loadings are statistically equivalent across countries, and the third test in hierarchy is that of scalar invariance (Rutkowski & Svetina, 2014).

However, most ILSAs fail to achieve full MI after conducting traditional statistical tests for MI and recently, modern approaches to the evaluation of MI have been developed to overcome the drawbacks from traditional approaches when implemented on

data from ILSAs. Therefore, more research is needed to (a) identify the limitations of traditional statistical approaches, (b) determine the extent to which those limitations impact the analysis when the data is collected through ILSAs, and (c) explore modern approaches and their effectiveness to overcome the limitations from the traditional approaches.

2.4 Statistical Approaches for the Evaluation of Measurement Invariance

As previously mentioned, the evaluation of MI is a requirement when using data from ILSAs given that both countries and examinees are culturally and linguistically diverse. Measurement invariance testing provides evidence about the extent to which a measurement instrument has the same measurement properties across diverse populations and most testing procedures of MI focus on the relationship between each test item and the overall test score (Fischer & Karl, 2019; Jak, 2014). Specifically, the evaluation of MI informs about the extent to which discrepancies in test scores across countries or cultural groups represent performance differences on the target latent construct being measured instead of construct irrelevant differences (Oliveri & von Davier, 2013).

Measurement comparability of test scores can be conducted at the item and test level. Analyses at the item level typically include differential item functioning (DIF) methods such as parametric and nonparametric item response theory (IRT) and ordinal logistic regression whereas analyses at the test level include exploratory and confirmatory factor analysis, and comparisons of test characteristic curves (Oliveri et al., 2012).

However, the statistical approaches to the evaluation of MI can be classified into two broad categories: traditional and modern approaches. The approaches vary in several ways such as the consideration of the structure of the data (i.e., nested, multilevel), their

capacity to handle more than two groups, the requirements in terms of the extent to which variability and uncertainty are allowed in the estimations, and the required assumptions among others (Lomazzi, 2018). Data from ILSAs have several features that should be carefully considered in the selection of a statistical approach to test for MI.

2.4.1 Traditional Statistical Approaches to Evaluate Measurement Invariance

Traditional statistical approaches to measurement invariance identify three broad levels of measurement invariance: (a) configural invariance, which is the starting point for further analyses and focuses on testing the equivalence of the factor structure across groups; (b) metric invariance that provides information on whether the measurement instrument has the same structure across groups by testing for the equivalence of factor loadings; and (c) scalar invariance that provides information about mean equivalence across groups by testing for the equivalence of intercepts (Fischer & Karl, 2019; Gorges et al., 2017; Meng et al., 2018).

Accordingly, the evaluation of measurement invariance has been conducted through procedures that test for multigroup equivalence and follow hierarchical steps starting with the development of a well-fitted baseline multigroup model where sets of parameters are ordered and tested for equality and the subsequent analyses -that typically involve factor loading regression paths and the factor covariances- are conducted in an increasingly restrictive fashion (Byrne, 2004; Byrne & van de Vijver, 2010).

The analysis involves a “model trimming” strategy where an initial model (configural invariance) is gradually restricted by adding cross-group equality constraints in an ordered sequence (metric and scalar invariance). Specifically, the procedure begins with the least restrictive model that is, the configural model where no equality constrains

are imposed and the focus is on the evaluation of the extent to which the number of factors and patterns of parameters that are freely estimated hold across groups thus, the same model of the hypothesized factorial structure is tested for each group. Configural models serve then as baseline models against which the subsequent tests for equivalence will be compared and once the configural model has been established, subsequent tests for equivalence can be conducted:

1. Testing for structural invariance. These tests focus on unobserved or latent variables specifically, the parameters of interest are the factor covariances to determine the extent to which the dimensionality of the target latent construct holds across groups or the extent to which a scale yields the expected dimensional structure (based on a specific theory) across groups (Byrne, 2004; Byrne & van de Vijver, 2010; Kline, 2016; Meng et al., 2018).
2. Testing for measurement invariance. These tests are focused on aspects of the observed variables only. Two tests that focus on the equality of factor loadings across groups can be conducted and they typically involve item intercepts and their associated errors.

    2.1 Tests for metric invariance: focus on the equivalence of factor loadings and to do so one group is chosen arbitrarily to be the reference group for which the parameters will be estimated freely. Then, factor loading estimates for the remaining groups are constrained equal to those of the reference group and under conditions of equivalence, these factor loading parameters remain constrained across subsequent tests for equivalence of additional parameters. Tests for invariant factor loadings are based on the analysis of covariance

structures where it is assumed that the observed variables are measured as deviations from their means.

2.2 Tests for scalar invariance: evaluate equivalence among intercepts. These tests are more restrictive when compared to tests for metric equivalence and involve the evaluation of equality of item intercepts based on mean and covariance structures that is, the moment matrix that includes sample means and covariances.

Most statistical techniques within the traditional approaches to the evaluation of MI incorporate this model trimming strategy to some extent and the most commonly used techniques are multidimensional scaling, exploratory factor analysis, confirmatory factor analysis, and multiple group confirmatory factor analysis (Fischer & Fontaine, 2011).

2.4.1.1 Multidimensional Scaling (MDS). This technique does not require the specification of an a priori test structure and in this sense, it is similar to exploratory factor analysis. However, in MDS data from multiple groups can be analyzed simultaneously and fitted to all groups to further detect structural differences by analyzing the group weights (Sireci et al., 2005).

MDS is typically used to represent the observed associations among items as distances between points in a geometrical representation of the true associations between items where large positive associations are represented by small distances and large negative associations by large distances. To conduct the analysis a similarity matrix must be created for each cultural group where Euclidean distances between standardized variables are inversely monotonically related to the Pearson correlations between the variables (Fischer & Fontaine, 2011).

The major weakness of MDS is that it is a descriptive technique that only generates an internal structure that best represents observed associations given a selected dimensionality and it does not provide statistical tests for observed structural differences among cultural groups (Fischer & Fontaine, 2011).

2.4.1.2 Exploratory Factor Analysis (EFA). EFA has been one of the most used approaches to test for construct equivalence and identify if the constructs have the same form and frequency across cultural groups (Sireci et al., 2005). EFA involves a measurement model that assumes that test items are indicators of unobserved latent constructs thus, the observed relationships among items are attributed to the latent constructs. This technique identifies the underlying constructs that maximally account for the common variance among test items; the input for the analysis consists of Pearson correlations among test items and the output is a matrix of factor loadings that includes correlations among test items and factors representing the latent construct (Fischer & Fontaine, 2011).

Typically, EFA appears to be more appropriate to represent the structure of a measurement instrument than confirmatory factor analysis (CFA) and this is because factor structures are not usually consistent with highly restrictive independent cluster models such as CFA where items are only allowed to load on a single factor and non-target loadings are constrained to zero (Marsh et al., 2011). The main disadvantage that has been associated to this technique in the context of the evaluation of MI is that the analyses are conducted separately for each culture group and the resulting factor loading matrices must be inspected individually (Sireci et al., 2005). Moreover, because of this limitation it is highly likely that this technique cannot accommodate the features that are

specific to data from ILSAs which could in turn compromise the validity of the interpretations to be made from these analyses.

2.4.1.3 Confirmatory Factor Analysis (CFA). CFA is a statistical modeling technique that incorporates latent variables as dimensions (factors) so that each indicator is allowed to depend on the factor(s) specified by the researcher and in this sense, CFA analyzes restricted measurement models. Factor-analytic methods aim to model the interrelations among indicators and to do so, the methods begin by partitioning the standardized variance into:

1. Common variance: shared variance among items that is assumed to be due to the factors. The proportion of shared variance is called communality.
2. Unique variance: consists of specific variance and measurement error. Specific variance is systematic variance that is not explained by the factors but could be the result of the features of individual indicators (Kline, 2016).

The main goal in CFA is to evaluate a theory-driven proposed structure of implied covariances among test items and compare it to the observed covariances that result from responses to the items. The quality of the models is evaluated through fit indices that can be grouped into two broad categories:

1. Incremental or comparative fit statistics that are typically used when a theoretical model is compared to an alternative model that does not include relationships among the variables. Higher values of the fit statistic suggest better fit, and the most common indices are the Tucker-Lewis Index (TLI), the non-normed fit index (NNFI) and the comparative fit index (CFI) (Fischer & Karl, 2019).

2. Lack of fit indices. These indices suggest better fit when the values are low and the most common include the standardized root mean square residual (SRMR) that compares the discrepancy between the observed correlation matrix and the implied theoretical matrix, and the root mean square error of approximation (RMSEA) that considers model complexity by rewarding the most parsimonious models (Fischer & Karl, 2019).

In terms of cutoff criteria for the indices, the values of RMSEA are expected to be close or below 0.060 while values of CFI and TLI are expected to be close or above 0.95 to indicate good model fit (Isac et al., 2019). However, values of RMSEA between 0.080 and 0.010 have been considered as indicators of acceptable fit and values of CFI and TLI between 0.9 and 0.95 are also considered as indicators of acceptable model fit (Isac et al., 2019).

CFA models are usually represented graphically through LISREL notation. See Kline (2016) for an example of a CFA model using this notation. CFA models must be identified to be analyzed. In short, measurement models are identified when it is possible to derive a unique estimate of every parameter in the model and in the context of standard CFA there are some general requirements for identification that must be met: (a) every latent variable in the model should be scaled, (b) the degrees of freedom of the model must be at least zero, (c) there should be at least three indicators for a single factor, and (d) models should have at least two factors (Kline, 2016).

Another important feature of CFA is that it can be used with unidimensional and multidimensional models. In the case of unidimensional models, indicators are assumed to be caused by the factor they are supposed to measure, and the error term represents all

unique sources of influence moreover, it is assumed that the error terms are independent of each other and of the factors. Multidimensional models on the other hand, typically incorporate complex indicators that are caused by two or more factors and can have at least one error correlation that represents shared sources of variation apart from the factors (Kline, 2016). Furthermore, CFA can also be implemented when the data are categorical such as dichotomous items, Likert-scale items, and partial credit polytomous items (Millsap & Yun-Tein, 2004).

Given these features, CFA has been widely used in the analysis of data from ILSAs. Specifically, CFA is commonly used to test for MI where several models are tested progressively, and constraints are added each time. In this sense, MI involves the evaluation of configural invariance where the patterns of factor loadings are examined to determine their equivalence across groups. The analyses typically begin with the assessment of metric invariance through a model where the factor loadings are set to be invariant across groups and then, restrictions are added on the intercepts to test for scalar invariance (Fischer & Karl, 2019; Rosen et al., 2013). However, one limitation of standard CFA is that it cannot incorporate several groups at the same time in the analysis therefore, it is not recommended when the data under analysis involves several groups as it is the case of data from ILSAs.

2.4.1.4 Multiple Group Confirmatory Factor Analysis (MGCFA). Multiple group confirmatory factor analysis (MGCFA) has been proposed as an alternative to traditional CFA where the analyses are conducted for one group at a time. MGCFA is used to simultaneously fit a model to data from multiple samples where group differences on parameters can be directly tested through the specification of cross-group equality

constraints and it assumes that there is a linear function between the indicator variables and the continuous target latent construct (Kline, 2016; van de Vijver et al., 2019).

This technique is based on a confirmatory factor analysis model for each group *j* with observed scores for individual *i* within group *j* so that:

$$Y_{ij} = \tau_j + \Lambda_{j\eta_{ij}} + \varepsilon_j \quad (2)$$

where, τ_j and Λ_j represent the intercepts and factor loadings, respectively while η_{ij} and ε_j represent the common factors and residuals, respectively. In this model, the observed score Y_{ij} is a linear function of the common factor score η_{ij} weighted by factor loadings Λ_j and the intercept τ_j (Kim et al., 2017). MGCFA is one of the most commonly used technique to test for MI by testing a sequence of measurement models from the least to the most restrictive in terms of the constraints that are placed on the measurement parameters across groups (Cieciuch et al., 2014; Davidov et al., 2018; Seddig & Lomazzi, 2019; Verdín & Godwin, 2017).

The analysis of MI in the context of MGCFA considers three hierarchical levels of measurement invariance:

1. Configural invariance, where there are no equality restrictions across groups that is,

   $$y_{ij} = \nu_j + \lambda_j f_{ij} + \epsilon_{ij} \; and \; E(f_j) = \alpha_j = 0, V(f_j) = \psi_j = 1 \quad (3)$$

   This type of invariance is achieved when the items within a measurement instrument exhibit the same pattern of factor loadings across groups suggesting that the dimensionality of the target latent constructs is the same across multiple groups.

2. Metric invariance, where the values of the factor loadings are assumed to be equal across groups making it possible to compare factor variances and structural relationships in structural equation modeling that is,

   $$y_{ij} = \nu_j + \lambda f_{ij} + \epsilon_{ij} \; and \; E(f_j) = \alpha_j = 0, V(f_j) = \psi_j \quad (4)$$

   This type of invariance takes place when the items show the same factor loadings across groups suggesting that the indicators are related to the target latent construct in the same way across groups.

3. Scalar invariance, where it is specified that the factor loadings and measurement intercepts are invariant across groups making it possible to compare factor means and factor intercepts that is,

   $$y_{ij} = \nu + \lambda f_{ij} + \epsilon_{ij} \; and \; E(f_j) = \alpha_j, V(f_j) = \psi_j \quad (5)$$

   This type of invariance is achieved when the items have the same intercepts or item difficulty across groups in addition to having the same discrimination and pattern of factor loadings suggesting that there are no differences in the average item responses across groups that are not due to differences in the mean level of the target latent construct.

For each of the models, i denotes an individual, j denotes a group, ν is the measurement intercept, λ represents a factor loading, f represents a factor with mean α and variance ψ, ϵ denotes a residual with mean zero, and variance θ that is uncorrelated with f. Moreover, the configural model has the subscript j for the intercepts and loadings, while the metric model does not include the subscript j for the loadings, and the scalar model does not include the subscript j for neither the intercepts nor the loadings. As shown, the configural model cannot identify a factor mean and variance because the

intercepts and loadings are set to be noninvariant however, it sets the metric of the factor by fixing the factor mean to zero and the factor variance to 1. The metric and scalar models on the other hand, identify group differences in the factor variances and in the factor means and variances, respectively (Fischer & Karl, 2019; He et al., 2019; Isac et al., 2019; Jak et al., 2014; Marsh et al., 2011; Marsh et al., 2018; Martin et al., 2019; Muthén & Asparouhov, 2018; Oliveri et al., 2012; van de Vijver et al., 2019).

In this context, scores cannot be compared across groups based on configural invariance alone; metric and scalar invariance must be met so that the former allows for the comparison of parameters that express relationships among construct while the latter is the only that allows for valid comparisons of latent means (Fischer & Karl, 2019; He et al., 2019; Isac et al., 2019; Jak et al., 2014; Marsh et al., 2011; Marsh et al., 2018; Martin et al., 2019; Oliveri et al., 2012; van de Vijver et al., 2019; van de Vijver & Leung, 2011; Verdín & Godwin, 2017). In this context, each level of invariance includes the previous and MI is assumed to be achieved if the more constrained model does not fit the data significantly worse than the less constrained model (Martin et al., 2019).

The models are typically evaluated through model fit indexes that assess the extent to which model fit deteriorates when moving from a configural to a metric model and from a metric to a scalar model. The most used measures of model fit are Comparative fit index (CFI), Tucker Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) and Standardized Root Mean Square Residual (SRMR). It is important to highlight that RMSEA tends to become greater than .05 regardless of model fit when the number of groups is large (e.g., more than 10 groups) thus, it is recommended to use an RMSEA cutoff of .10 when evaluating several groups. In

general, the following criteria for model fit indexes is suggested when comparing large number of groups: ΔCFI≤ .02 and ΔRMSEA ≤.03 for metric invariance and ΔCFI≤ .01 and ΔRMSEA ≤.015 for scalar invariance (Fischer & Karl, 2019; Isac et al., 2019; Kim et al., 2017; van de Vijver et al., 2019).

In this regard, Rutkowski and Svetina (2017) conducted a simulation study to determine if the traditional accepted measures for evaluating measurement invariance are suitable for large numbers of groups and non-normal observed variables from cross-cultural surveys. They simulated categorical data and generated parameters based on empirical results from the Teaching and Learning International Survey (TALIS). They conducted hierarchical tests and fitted MGCFA models for 23 countries starting with a baseline model that was followed by increasingly restrictive tests of equal slopes and equal slopes and thresholds. Based on the results, the authors made the following recommendations regarding model fit indices when testing for measurement invariance in a multiple-group context:

1. Overall test: cutoff of .055 for RMSEA when testing for configural, metric and scalar invariance.
2. Incremental tests: (a) metric invariance: a cutoff of -0.004 for the change in CFI and a cutoff of .05 for the change in RMSEA, and (b) scalar invariance: a cutoff of -0.004 for the change in CFI and a cutoff of .01 for the change in RMSEA.

Moreover, they advised that neither the CFI nor the TLI indices should be used to evaluate the overall fit of multiple-groups models with different sample sizes.

On the other hand, MGCFA can also be applied when the data are categorical as it is the case for most non-cognitive measures. The multiple-group factor model for

categorical measures can be stated letting X_{ijk} be the score on the j^{th} ordered-categorical measure for the i^{th} person in the k^{th} group. In most cases, all measured variables have score ranges $\{0,1,\dots,c\}$ where *c* represents the largest possible score, and the number of measured variables is represented by $p(j = 1,\dots,p)$. The factor model for categorical data assumes that observed scores X_{ijk} are determined by unobserved scores on the latent response variables X^*_{ijk} which are continuous thus the observed variables can be seen as discrete versions of the latent response variables since scores on observed variables are given by

$$X_{ijk} = m \quad if \quad v_{jkm} \leq X^*_{ijk} < v_{ij(m+1)} \tag{6}$$

where $m = 0,1,\dots,c$ and $\{v_{jk0}, v_{jk1},\dots, v_{jk(c+1)}\}$ are the latent thresholds parameters for the j^{th} variable on individuals from the k^{th}group. Two of the thresholds are predefined: $v_{jk0} = -\infty$ and $v_{jk(c+1)} = +\infty$ and the remaining *c* thresholds parameters can vary across variables and groups. Moreover, the probabilities of the observed values of X_{ijk} are given by the probability distribution of X^*_{ijk}. Thus, if $X_{ik}{}' = \{X_{i1k}, X_{i2k},\dots, X_{ipk}\}$ is the $1 \times p$ vector containing observed scores on the *p* variables for the i^{th} person from the k^{th} group with $X_{ik}{}'$ the analogous vector of scores on the latent response variables, then it is assumed that

$$X_{ik} \sim MVN(\mu^*_k, \Sigma^*_k) \tag{7}$$

where μ^*_k is a $p \times 1$ vector of means on the latent variables and Σ^*_k corresponds to the $p \times p$ covariance matrix for the latent variables. These parameters are allowed to vary across groups (Millsap & Yun-Tein, 2004; Rutkowski & Svetina, 2017).

Empirical evidence about the suitability of MGCFA to test for MI has been mixed. For instance, Rikoon and Midkiff (2018) conducted a study to evaluate the

measurement invariance of the SuccessNavigator ® assessment across three cohorts of undergraduate students. The authors built longitudinal factor models that specified one latent variable at each time point representing the subskill of interest. In short, the authors first specified a baseline model that only included identification constraints and then, that model was modified in stages by adding series of constraints with different sets of parameters constant over time. Invariance models were specified in three stages: (a) a first stage where the configural model was used where the same pattern of associations among latent variables and observed indicators were specified at each time point and all loadings and intercepts were freely estimated meaning that this model specified the same measurement structure for a subskill at each time point but allowed the relationships between the target latent variables and their observed indicators to vary over time, (b) in the second stage a model for each subskill specifying metric longitudinal measurement invariance was estimated and the metric model was specified so that the factor loadings for each observed indicator were constrained to be the same across time points and this model was compared to the configural model, and (c) in the third stage constraints on item intercept parameters over time were added into the a scalar invariance model, which was compared to the metric model. The authors found evidence suggesting that eight out of the 10 subskills showed partial scalar measurement invariance so that the factor structure and psychometric properties of the items within the scale functioned similarly over time which supported the comparison of longitudinal mean level changes within the sample.

In a similar way, Melendez-Torres et al. (2019) found evidence of MI through the analysis of MGCFA. They conducted a study to assess the psychometric properties of a

wellbeing measure that was developed to target adult population and tested for measurement invariance when a shorten version of the measure was administered to adolescent population. First, they tested for configural invariance by estimating polychoric correlation matrices for the whole sample and conducted principal component analysis to determine if the number of factors was equal across year groups. Then, they used MGCFA models with successively greater constraints to test for measurement invariance and since the data were categorical, they used a diagonally weighted least square estimator with a scale-shifted test statistic. Regarding the models, the authors began with a first model assuming configural invariance, the second model restricted factor loadings to be equal across groups, the third model additionally restricted thresholds to be equal across groups, and the fourth model set the residual variance to be equal across groups. To evaluate model fit, the authors used the comparative fit index (CFI) and the root mean squared error of approximation (RMSEA). The authors found evidence suggesting that loadings and thresholds were invariant.

On the other hand, Verdín, and Godwin (2017) conducted a study to test for measurement invariance between first generation and non-first-generation college students on a measure of engineering identity. They conducted the analyses in a stepwise fashion where they first tested for configural invariance by evaluating a three-factor structure model for the latent construct. In this phase, the authors did not place any equality constraints so that all parameters were freely estimated for each group separately. After evaluating the fit indexes, they found evidence of configural invariance and thus, found a basis for conducting a multiple group CFA to test for model invariance. Once they collected evidence of configural invariance, they proceeded to test for metric

invariance where they constrained the factor loadings to be equal across groups and evaluated model fit. They found evidence of metric invariance and concluded that the factor loadings could be estimated simultaneously and that the model had the same structure across groups. Then, they tested for scalar invariance but found no evidence for this level of measurement invariance thus, they concluded that the groups responded differently to the items from the scale and thus, further comparisons of the composite scores on the constructs will be biased.

Similarly, He et al. (2019) evaluated the measurement invariance of the motivation, sense of belonging to school and enjoyment of science scales within the PISA 2015 student questionnaire. They used MGCFA and they treated the data as continuous. For the analysis they implemented the full information maximum likelihood estimation method and used the senate weights that rescale sample sizes to be fixed at 500 cases per country. According to the authors the use of senate weights is recommended to balance the contribution of each country in the estimation. The model fit in MGCFA was assessed through the Chi-square statistic, the comparative fit index (CFI) and the root mean square error of approximation (RMSEA) while the acceptance of a more restrictive model was based on the change of CFI and RMSEA using a cutoff of .02 for the change in CFI and .03 for the change in RMSEA when evaluating the change from configural to metric models, and a cutoff of .01 for the change in both indices when evaluating the change from metric to scalar models. They found that the motivation scale showed metric invariance, the enjoyment of science scale showed configural invariance and acceptable metric invariance while the configural model for the scale of sense of belonging to school did not converge. Scalar invariance was not found for any of the scales.

The fact that empirical findings are not consistent and that some studies report evidence of full invariance whereas others do not find evidence to support invariance could be due to limitations that have been highlighted regarding the application of MGCFA techniques for the assessment of MI; specially, when the analyses involve more than two groups and when the data have been collected through ILSAs. According to van de Vijver et al. (2019), the major limitation of MGCFA is that it is too strict since it requires exact equality of parameters across groups which is not suitable to achieve in real data analysis especially when sample sizes and the number of groups are large as it is the case of ILSAs.

In a similar way, Rutkowski and Svetina (2014) pointed out that even though MGCFA is one of the most commonly used techniques to test for MI on data from ILSAs, evidence has shown that this technique might not be appropriate to handle large-scale data from international assessments in terms of its performance and resulting fit indices when applied to large sample sizes. Moreover, Hox et al. (2012) found that when the number of countries is large the approach of MGCFA is unmanageable and this is related to the fact that MGCFA is a fixed effects model that estimates a unique set of parameter values for each country.

Asparouhov and Muthén (2014) also pointed out the number of groups analyzed is a major limitation of MGCFA when applied to data from ILSAs. As they stated, “With many groups, the usual multiple-group CFA approach is too cumbersome to be practical due to the many possible violations of invariance, and the modification index exploration could well lead to the wrong model due to the scalar model being far from the true model” (p. 1).

The challenge with the analysis of large number of groups has to do with the fact that the number of pairwise comparisons across groups on measurement parameters increases as the number of groups increase; and as the number of pairwise comparisons increase so do the chances of falsely detecting noninvariance. Additionally, model fit criteria are often too stringent when the number of groups is large (Asparouhov & Muthén, 2014; Byrne & van de Vijver, 2017; Kim et al., 2017). The fact that some studies do not report any limitations with the technique is probably because in most cases the comparisons are conducted only between two groups, which may not be appropriate when analyzing data from ILSAs where several countries are being compared, and for which measurement invariance is expected to hold despite their cultural, language, and geographical differences (Rutkowski & Svetina, 2017).

Other limitations in using MGCFA to evaluate invariance include (a) the assumptions about the equality of several parameters across a large number of groups is highly implausible when analyzing real data, (b) scalar invariance is rarely achieved when tested on data from ILSAs, (c) stepwise approaches to measurement invariance rely on modification indices to make post hoc corrections that are typically problematic due to the violations of statistical estimation and hypothesis testing that the procedures involve, and (d) data are expected to be normally distributed, an assumption that is not likely to hold in ordinal data from non-cognitive measures (Fischer & Karl, 2019; Marsh et al., 2018; Rutkowski & Svetina, 2017).

Given that in the context of MGCFA it is difficult to achieve full measurement invariance for international large-scale data, some researchers and test developers have decided to report partial measurement invariance and have placed a greater emphasis on

metric invariance. However, making a statement about measurement invariance based only on evidence of metric invariance does not seem to be the best approach because it is not possible to state that a test item is invariant and therefore, perceived in the same way by test takers if only the slope (factor loading) but not the intercept is invariant when the item is regressed on a factor (Muthén & Asparouhov, 2018).

In conclusion, traditional statistical approaches to the evaluation of MI have been widely used and among them, MGCFA has been the most popular. However, even though the techniques are straightforward, empirical evidence has highlighted several limitations when applied to data from ILSAs. Alternative statistical approaches that are suitable to handle the specific features of data from ILSAs need to be explored and evaluated to ensure the inferences about the extent to which the measures are invariant across countries and examinees are valid and thus, suitable to be used by governments throughout the world to inform their educational policies and practices.

2.4.2 Hierarchical and Latent-Based Statistical Approaches

Despite the popularity of traditional approaches to the evaluation of measurement invariance, they are not suitable to handle international large-scale data involving diverse cultural groups and the major limitations include: (a) it is not always possible to formulate a baseline model that is the same for all the groups, (b) in practice, the condition that all non-target factor loadings are fixed to zero across groups does not hold resulting in poorly fitting models and a large number of parameters that are not specified, and (c) tests for equality of constrained parameters are typically done by comparing two groups at a time making it difficult, if not impossible, to test large number of groups as it is the case when analyzing data from ILSAs (Byrne & van de Vijver, 2010; Byrne & van

de Vijver, 2017). New approaches to the assessment of measurement invariance have been developed recently to overcome the limitations of traditional techniques and most of them allow for tests of approximate measurement invariance such as exploratory structural equation modeling (ESEM), the alignment optimization method, and Bayesian structural equation modeling (BSEM) (Byrne & van de Vijver, 2017).

2.4.2.1 Multilevel Confirmatory Factor Analysis (MLCFA). Multilevel models have gained popularity in the recent years and have been largely applied in cross-cultural research where one of the main goals is to determine the extent to which the observed relationships among variables can be generalized across countries. To do so, the multilevel nature of cross-cultural data as that collected through ILSAs, where individuals (micro-level) are clustered within territorial units (countries at the macro-level), must be considered since the observed data at the macro level are impacted by the mechanisms at the micro level (Kim et al., 2016; Meuleman, 2019).

MLCFA has been often used to analyze complex survey data including data from ILSAs, through the estimation of level-specific variance components within the measurement models. Complex survey data are typically obtained through cluster sampling that results in non-independent observations with within-cluster dependency and thus, cannot be analyzed using traditional approaches that heavily rely on assumptions of independence (Wu et al., 2017).

The multilevel approach is simpler than the MGCFA since only one measurement model is constructed for all groups and it can also be implemented when there are many groups under analysis (Kim et al., 2017). Multilevel systems in general, allow for the modeling of variables at the between and within levels while characterizing the

relationship between unobserved latent factors and observed indicators. For instance, MLCFA is a multilevel analytical tool that decomposes the total sample variance-covariance matrix into within-cluster (typically the individual level) and between-cluster (e.g., country level) matrices while modeling distinct latent factor structures at each of the levels simultaneously. By doing so, it is possible to draw more accurate inferences about the performance of a set of items at both levels while understanding the meaning of the latent construct at each level of analysis (Dunn et al., 2015; Kim et al., 2016).

The modeling of multilevel data across multiple populations begins with a data structure where y_{gci} denotes a vector of variables for a randomly sampled individual *i* from cluster *c* for group g. This vector can be decomposed into between and within cluster variation:

$$y_{gci} = yB_{gc} + yW_{gi} \tag{8}$$

and

$$E(y_{gci}) = \mu_{yg} \tag{9}$$

Also, the total covariance matrix can be decomposed into a within and between cluster part:

$$\sum T_g = \sum B_g + \sum W_g \tag{10}$$

The latent variable model has a conventional factor analytic structure for the between and within cluster level. Thus, the between level is specified as

$$yB_{gc} = v_g + \Lambda B_g \eta B_{gc} + \epsilon B_{gc} \tag{11}$$

where v_g denotes the intercept parameter vector, ΛB denotes the between-level loading parameter matrix, ηB corresponds to the latent between-level variable vector, and ϵB denotes the between-level residual vector. Moreover,

$$E(\eta B_{gc}) = \alpha_g, \tag{12}$$

$$V(\eta B_{gc}) = \Psi B_g, \tag{13}$$

$$V(\epsilon B_{gc}) = \Theta B_g \tag{14}$$

* $E = expected$ and $V = variance$

and the within-cluster level is specified as:

$$yW_{gi} = \Lambda \mathrm{W}_g \eta W_{gci} + \epsilon W_{gci} \tag{15}$$

where yW_{gi} is the within-cluster variation with mean zero so that the intercept vector is zero and

$$E(\eta W_{gci}) = 0 \tag{16}$$

As for the mean structure, it is specified as

$$E(yW_{gi}) = 0 \text{ whereas } E(yB_{gc}) = v_g + \Lambda B_g + \alpha_g \tag{17}$$

so that the means appear only at the between level. The means are specified for the level of variation for which there are independent observations available, in this case the between level (Muthén et al., 1997).

In this approach groups are considered randomly selected from the population thus, instead of constructing one model for each group, a single measurement model representing the average model across the random groups is constructed with a pooled within-group covariance matrix and a between-group covariance matrix based on the randomly varying cluster means of the observed variables (Kim et al., 2017).

To test for measurement invariance, a configural invariance model is first developed where the same factor structures are specified at the within and between level, then the metric invariance is tested across groups or clusters by imposing a cross-level invariance constraint so that if factor loadings are the same across groups, the factor loadings of the within-group CFA model should be identical to those of the between-group CFA model. Also, if the intercepts are the same (not random) across all groups, the between-group variability of intercepts should be equal to zero and this scalar invariance is tested by constraining the between-level residual variance to zero (Kim et al., 2016; Kim et al., 2017).

The evaluation of measurement invariance begins by identifying a reasonable factor structure for the within and between-group measurement models, then configural invariance is tested by constructing the same factor structure for both models and the factor loadings are allowed to vary across levels except for the factor loading of the first item that is fixed to 1 at both levels for model identification purposes, while the residual variances at the between level are freely estimated. Then, if configural variance holds, scalar invariance is tested by constraining the factor loadings of the within-group measurement model to be the same as those of the between-group measurement model. Evidence of scalar invariance is achieved if the scalar invariance model is selected over the configural invariance model (Kim et al., 2017).

On the other hand, two major limitations have been associated to the use of MLCFA for the assessment of MI form ILSAs: (a) it only tests the equivalence of item discrimination across groups and does not consider item difficulty parameters, and (b) the achievement of complete scalar invariance based on this approach is not feasible

specially in the context of ILSAs; in this scenario, only approximated scalar invariance would be feasible and yet, acceptable approximations to complete scalar CFA-MI are still rare in international large-scale studies featuring large numbers of groups, factors and items (Marsh et al., 2018; Oishi, 2006).

In conclusion, MLCFA has been suggested as suitable alternative to MGCFA and its main advantage is that it allows for the incorporation of the multilevel structure of the data into the estimations. This feature is particularly relevant when the data under analysis has been collected through ILSAs where the data is nested in terms of the group membership of examinees (i.e., cultural, linguistic background) and of the countries.

However, some empirical evidence suggests that this technique does not allow for the achievement of scalar invariance with data from ILSAs given the differences among groups in terms of sample size and the large number of groups under analysis thus, it has been suggested that approaches based on approximation to measurement invariance could be a suitable alternative to handle the features of data from ILSAs. More evidence should be collected to evaluate the extent to which the estimations from MLCFA lead to accurate inferences from test scores in the context of ILSAs.

2.4.2.2 Multilevel Structural Equation Modeling (MSEM). As previously mentioned, ILSAs collect data on individuals across countries and in order to make proper inferences about the test takers, their culture-specific features must be acknowledged given the well-known influence that culture has on the way people think, behave, and communicate (Cheung et al., 2006). Moreover, a common feature in ILSAs is the use of cluster sampling where higher-level units are randomly selected, and then lower level units are selected within the higher-level ones thus, two levels of analysis are

easily identified in data from ILSAs: individual and country/culture, and this nonindependence of individuals within cultures must be properly addressed when modeling these type of data through multilevel models that account for the individual and culture levels simultaneously by differentiating between (a) within-group variances and effects that are related to deviations from the mean, and (b) between-group variances and effects that are related to group means (Cheung et al., 2006; Jak et al., 2014; Zigler & Ye, 2019; Zyphur et al., 2019).

When the multilevel structure of the data is not considered in the analyses, biased statistical inferences are likely to occur for instance, standard errors are likely to be underestimated because data from the same culture will be more similar than data across cultures which in turn, increases the likelihood of type I error. Moreover, statistical analyses that do not consider the hierarchical structure of data can yield misleading results particularly when the results from the group level are interpreted at the individual level or vice versa (Cheung et al., 2006; Christ et al., 2017).

In the past decades, structural equation modeling (SEM) has been the preferred statistical modeling technique to test for latent mean differences across groups and therefore, to test for measurement invariance. SEM is a causal inference method based on (a) either a set of qualitative causal hypothesis based on theory or results of empirical studies that are represented in a structural equation model, and (b) a set of questions about causal relations among variables (Byrne & van de Vijver, 2017; Kline, 2016). The primary input in SEM analysis is the covariance which can be defined for two continuous variables as:

$$cov_{XY} = r_{XY} SD_X SD_Y \tag{18}$$

where r_{XY} is the Pearson correlation and SD_X and SD_Y are the standard deviations thus, the covariance represents the strength of the linear association between X and Y plus their variabilities. Given that the covariance is the main statistic in SEM, two major goals can be identified for the SEM analysis: (a) understand patterns of covariances among a set of variables, and (b) explain as much of their variance as possible with the proposed model. In SEM analysis the part of the structural equation model that represents the hypothesis about variances and covariances is referred to as the covariance structure (Kline, 2016).

In the context of SEM, the hypotheses of interest are usually first depicted as graphical conceptual models that are eventually translated into statistical models described by a series of equations that define model parameters which correspond to the assumed relations among variables. The major requirement for statistical models within the SEM framework is identification and models are identified if it is theoretically possible to derive a unique estimate of every model parameter (Kline, 2016).

Standard SEM models the covariance and mean structures of multivariate data by estimating parameters that reproduce observed data structures. SEM includes both measurement models to estimate the relationships between observed indicators and latent constructs, and structural models to estimate relationships among latent constructs. In this context, measurement models can be defined by:

$$Y_i = \nu + \Lambda\eta_i + \varepsilon_i \qquad (19)$$

where Y_i is the vector of observed responses to the indicators, η_i denotes latent variables for each individual (i), Λ denotes the matrix of factor loadings, ν represents the vector of item intercepts, and ε_i represents the residuals.

The structural models on the other hand can be defined by:

$$\eta_i = \alpha + \mathrm{B}\eta_i + \Gamma X_i + \zeta_i \quad (20)$$

where B denotes the matrix of effects among latent variables, Γ denotes the direct effects of all the exogenous variables ($X's$) (i.e., country, culture) on the latent variables, α represents the intercepts and ζ_i refer to the residuals of the endogenous variables. It is notable that the sets of estimates for both the measurement and structural parameters imply a covariance matrix (Σ) and mean structure and the most suitable set of parameters is obtained by minimizing the difference between observed means and covariances and the model-implied means and covariances (Meuleman, 2019).

When the data under analysis have a multilevel structure as in the case of data from ILSAs, the evaluation of measurement invariance can be challenging mostly because the standard SEM approaches need to be adjusted to consider the multilevel structure (Byrne & van de Vijver, 2017; Jak et al., 2014). Empirical evidence has shown that in educational settings, the factors at the between level such as culture and classroom climate can have an influential role in academic achievement and thus, need to be considered within this multilevel perspective. Considering the hierarchical nature of the data where individuals cannot be detached from their broader social context is crucial especially in terms of construct validity because a target latent construct can have a different operational meaning across levels of analysis which in turn, influences the interpretations made based on test scores (Christ et al., 2017; Sideridis et al., 2018).

Multilevel structural equation modeling (MSEM) is a modern statistical approach that includes features to allow for the modeling of multilevel data. The technique has gained popularity in cross-cultural research because it incorporates a latent-variable approach into the multilevel framework and therefore, combines the advantages of SEM

and multi-level modeling (MLM). In doing so, MSEM allows for the modeling of variances and covariances for within and between group differences by decomposing the data into between and within-group components that are orthogonal and additive (Byrne et al., 2009; Cheung & Au, 2005; Christ et al., 2017; Hox et al., 2012; Jak et al., 2014; Meuleman, 2019; Sideridis et al., 2018; Zigler & Ye, 2019). Latent variable modeling of multilevel data through SEM has been used in educational settings especially when students are sampled within clusters in large-scale assessments. Given that in most cases the clusters vary in several characteristics, it cannot be assumed they were sampled from a single common population. Therefore, the aim is to generalize the mean and covariance structure modeling of multilevel data to the analysis of multiple populations (Muthén et al., 1997).

MSEM assumes that the population covariance matrices are described by different models for the between and within groups structure providing within and between-level parameters to describe the structure of within-group variables (e.g., differences across individuals within countries) and the relationships among the between-level variables (e.g., group averages of country-level variables) (Hox et al., 2012; Jak et al., 2014; Meuleman, 2019).

In more detail, MSEM considers that individuals (i) from a population can be hierarchically nested within groups (g) which are usually countries thus, MSEM allows for the orthogonal decomposition of observed scores into (a) a group or between component such as the group average and, (b) an individual or within component such as the deviation from the group average so that:

$$y_{ig} = \overline{y_g} + (y_{ig} - \overline{y_g}) \quad (21)$$

where $\overline{y_g}$ represents the group average and $y_{ig} - \overline{y_g}$ the deviation from the group average.

Based on this decomposition the total covariance structure will be split into two covariance matrices:

$$\Sigma_T = \Sigma_B + \Sigma_W \quad (22)$$

where Σ_W is the within covariance structure that summarizes how the individual components are related, Σ_B denotes the between covariance structure that describes how the group level components of the variables covary. Therefore, MSEM estimates separate effects for the within and between level components of individual variables so that the within-level effects must be interpreted in terms of the differences among individuals within the groups (Jak et al., 2014; Meuleman, 2019).

In terms of the multilevel component, MSEM allows for the simultaneous modeling of the covariance structures for the within and between levels. Therefore, within and between models are formulated to reproduce the hierarchical structure of the data.

The discrepancies between the data and hypothesized variance-covariance matrices are typically evaluated through an omnibus chi-square test using a system of linear equations and model fit is usually assessed using descriptive fit indices and residual values including: (a) the comparative fit index (CFI), a goodness of fit statistic that can take values from 0 to 1.0 where 1.0 represents best fit and compares the amount of departure from close fit for the researcher's model against that of the null model, (b) the Tucker-Lewis index (TLI) that controls for df_M from the researcher´s model and the df_B from the baseline model while imposing a greater penalty for model complexity than

the CFI, (c) SRMR, an absolute fit index that measures the average squared covariance residual and informs about the overall difference between observed and predicted correlations, and (d) the RMSEA, an absolute fit index scaled as a badness of fit index so that a value of zero indicates the best fit and it measures departure from close or approximate fit. This index is sensitive to violations of normality and models with few variables (Kline, 2016; Sideridis et al., 2018).

Given the doubts about the trustworthiness of thresholds for some fit indexes, Kline (2016) suggested an alternate approach to model fit evaluation that focuses on reporting more specific information about model fit. Suggestions included:

- For simultaneous estimation methods, report chi-square with its degrees of freedom and associated *p*-value. Then, diagnose the magnitude and possible sources of misfit through local fit testing to detect model-data discrepancies that even though might not be statistically significant, could raise questions about the model.
- Report the matrix of residuals and describe their pattern. When inspecting model fit it is important to evaluate correlation residuals so that absolute correlation residuals higher than 1 could suggest poor local fit.
- When reporting approximate fit indexes, avoid making claims based solely on them.
- In general, always report at least: (a) chi-square with associated degrees of freedom and *p*-values, (b) root mean square error of approximation with its 90% confidence interval, (c) Bentler comparative fit index, and (d) standardized root mean square residual.

On the other hand, SEM in general is considered a key methodological approach to test for measurement invariance and MSEM in particular, has been recently used to test for measurement invariance across cultural groups (Byrne et al., 2009; Lee et al., 2017).

The evaluation of measurement invariance also involves the evaluation of configural, metric and scalar invariance. As previously mentioned, configural invariance suggests that the target latent constructs being measured are comparable across cultures and in the context of MSEM this similarity of meaning is needed at the individual and cultural levels to interpret the aggregated means with the same meaning as those in the individual level that is, the factor structures should be similar at both levels (Cheung et al., 2006).

Metric invariance is achieved when the factor loadings are equal across levels indicating that the common factor has the same meaning at the within and between levels, and scalar invariance is achieved when in addition to metric invariance, the residual variance at the between-level equals zero suggesting that the observed differences between groups are assumed to be due to differences in the common factor. Moreover, in MSEM it is possible to add between-level variables to explain the differences in the common factor as well as group-specific differences in specific items (Seddig & Lomazzi, 2019).

In terms of the estimators, the most common approaches to estimate the parameters in MSEM include an approximation of the full information maximum likelihood estimator and the weighted least squares method (Hox et al., 2012). Among these, the Full information maximum likelihood (FIML) is the most used estimator to

analyze multilevel data due to its optimality to lead to the smallest possible standard errors. However, FIML can be computationally demanding specially when applied to unbalanced data and model specifications can also be tedious which is why a simpler estimator has been proposed: the Muthen's maximum likelihood-based estimator (MUML) that leads to similar results as those obtained through FIML but with rough approximations to the correct chi-square test statistics and standard errors associated to parameter estimates (Cheung & Au, 2005). In the case of categorical variables, there are three alternate estimators: (a) the fully weighted least squares (WLS) estimator that does not assume a particular distributional form, (b) the robust WLS estimation which uses simpler matrix calculations than the full WLS and generates corrected standard errors and model test statistics; and (c) a version of the FIML that relies on numerical integration to estimate response probabilities in joint multivariate distributions of latent variables assumed to underlie observed categorical data (Kline, 2016).

Among these, the most used estimator is the robust weighted least square (WLS) estimation that does not make distributional assumptions and requires large sample sizes. In this context, each ordinal indicator is associated with a latent response variable representing the underlying amount of a continuous and normally distributed continuum required to respond on the indicator. Polytomous items have several thresholds or points on the latent variable where the response option equals the number of categories minus one. The logic behind the analysis of categorical items is described next.

If an item *X* has three response categories, the scale for the response is a categorization of *X** that is, the underlying latent response variable and in this case the item has two threshold parameters τ_1 and τ_2. When *X** has a mean of 0 and variance of 1

the thresholds represent values of the normal deviate (z) that divides the normal distribution into categories relating the discrete responses on *X* to continuous *X** values so that the data is represented as:

$$X = \begin{cases} 1, & if\ X^* \leq \tau_1; \\ 2, & if\ \tau_1 < X^* \leq \tau_2; \\ 3, & if\ X^* > \tau_2 \end{cases} \quad (23)$$

where a response of 1 is expected if the level of *X** is less than that of τ_1 in standard deviation units while for levels of *X** greater than τ_1 but less than or equal to τ_2, the expected response is 2, and finally, when *X**$> \tau_2$, the expected response is 3. The thresholds are estimated based on cumulative response probabilities and for a set of items, the estimated thresholds and observed cross tabulations of item response are used to estimate the matrix of Pearson correlations between the latent response variables that corresponds to the polychoric correlation. An asymptotic covariance matrix of the polychoric correlations is generated whose inverse is the weight matrix in full WLS estimation. The diagonal elements of the asymptotic covariance matrix estimate the variance of the polychoric correlations over random samples while the off-diagonal elements represent the covariances between the estimates moreover; robust WLS estimation uses the diagonal of the asymptotic covariance matrix in its fit function. In the analysis, the relations between the latent response variables and indicators are nonlinear and the parameters estimated through robust WLS are derived so that the correspondence between the observed polychoric correlations and those predicted by the model is as close as possible (Kline, 2016).

When analyzing categorical data, measurement invariance for an individual item means that the probability of selecting a response option is the same across groups given

the same level on the common factor that corresponds to that item. This property should hold for all items under analysis in order to establish measurement invariance (Kline, 2016).

For identification purposes of models involving categorical variables it is suggested that (a) the residual variance of each X* variable is fixed to 1 in a group that is designated as the reference group, (b) the mean of the common factor is zero in the reference group and the variance of every residual is standardized, (c) the direct effect of the constant on every X* is fixed to zero in every group and then the same X* is selected across the groups as the reference variable to fix its unstandardized pattern coefficient to 1, (d) one threshold parameter is constrained to equality across groups for every X*, (e) one threshold parameter is constrained to equality across groups for every X* that is a reference variable (Kline, 2016).

On the other hand, despite the advantages that have been associated to MSEM, empirical evidence has shown that it also has some limitations that are worth mentioning. For instance, according to Cheung and Au (2005), the main drawback of the implementation of MSEM has to do with the minimum required sample sizes at the individual and group levels. It has been suggested that the MUML estimator performs well when the group-level sample size is of at least 100 which is a difficult requirement to meet in cross-cultural research where the number of groups typically ranges from 20 to 30.

In a similar way, Christ et al. (2017) pointed that a critical issue with the implementation of MSEM had to do with the sample sizes that are needed to obtain unbiased estimates of parameters and sufficient statistical power since the ideal sample

size at the between (group) level is 100, a sample size that is in practice difficult to obtain. However, it has been suggested that this limitation can be partially addressed with the implementation of Bayesian estimation methods.

Bayesian estimation combines prior knowledge with observed evidence about the likelihood of data given a set of parameters, to generate a posterior distribution of parameter estimates that expresses the level of uncertainty about the parameters that is left after having observed the data. Posterior distributions are obtained through an iterative procedure therefore, Bayesian estimation does not make distributional assumptions of test statistics neither does depend on the large-sample theory (Meuleman, 2019).

Another limitation pointed by Cheung et al. (2006) has to do with the fact that MSEM assumes that the proposed within-structure model at the individual level is the same across groups or cultures, an assumption that is rarely met in cross-cultural research where some psychological processes are not universal across cultures therefore, if this assumption is not met, results from MSEM can be misleading. On the other hand, Marsh et al. (2018) pointed out MSEM has been found to perform better when there are a large number of indicators as opposed to situations where the number of items is small as it is the case of measures of non-cognitive constructs.

In conclusion, MSEM has been proposed as a suitable approach to test for measurement invariance in data collected through ILSAs given that it accounts for the nested structure of the data and in this sense, it addresses the limitations associated to the traditional approaches to measurement invariance and to MGCFA. However, as mentioned, several limitations have been associated to this modeling technique therefore,

sound empirical evidence should be collected to determine the extent to which the approach is suitable to handle data from ILSAs.

2.4.2.3 Alignment Optimization. Traditional approaches to test for measurement invariance are known for their limitations when applied to international large-scale data mostly because as the number of groups increases so does the likelihood of not meeting the requirements for the establishment of full invariance. In this context, the alignment method has been suggested as a suitable alternative to test for measurement invariance that can be particularly useful when the data requires the analysis of several groups as in the case of data from ILSAs (Lomazzi, 2018).

The alignment optimization integrates the IRT and SEM approaches in the search of an optimal pattern of measurement invariance across many groups by allowing for a certain amount of non-invariance. Thus, the factor means can be estimated without equality constrains on loadings and intercepts across the groups because the goal is to keep the number of noninvariant parameters and the level of non-invariance to a minimum (Marsh et al., 2018; Munck et al., 2018; Muthén & Asparouhov, 2014; Seddig & Lomazzi, 2019).

The most salient advantages of the alignment optimization include (a) it can estimate models for several groups, (b) it automates the analyses by considering the non-invariance of all factor loadings and intercept parameters in the process of mean estimation resulting in trustworthy mean values despite the presence of some measurement noninvariance, (c) it simplifies the tests for measurement invariance across a large number of groups, (d) it provides a detailed account of parameter invariance for every model parameter within each group, and (e) it can handle the complex features of

data from PISA where sampling weights are used because the sampling procedure to select the schools was based on probability proportional to size (Byrne & van de Vijver, 2017; Fischer & Karl, 2019; Lamm et al., 2019; Muthén & Asparouhov, 2013; Muthén & Asparouhov, 2014; Muthén & Asparouhov, 2018; van de Vijver et al., 2019).

In this sense, the method allows for the comparison of several groups within countries by identifying a solution with the least measurement noninvariance and the best possible fit among all possible multiple-group CFA models. It also allows for the identification of groups that deviate from the common measurement pattern while detecting invariance for every model parameter within each group (Munck et al., 2018).

The implementation involves two general steps: (a) fitting a configural model across the groups where the loadings and intercepts are freely estimated while the factor means and factor variances are fixed at zero and one, respectively, and (b) free estimation of factor means and variances through the implementation of a simplicity function that minimizes the total amount of noninvariance across all model parameters. The analysis is repeated using different starting values to find an optimal and stable solution across iterations so that the final solution includes the fit function contribution of each item parameter to the alignment simplicity function across groups while detecting sources of noninvariance in the parameter estimates in a common metric (Fischer & Karl, 2019; Marsh et al., 2018; Munck et al., 2018).

By doing so, it is possible to estimate the simplest model with the largest amount of invariance where the quality of the alignment solution will depend on the presence of a minority of measurement parameters in the grouping that carry noninvariance which according to simulation studies should be no more than 25% (Munck et al., 2018).

In more detail, the method begins by considering the multiple-group factor analysis model:

$$y_{ipg} = v_{pg} + \lambda_{pg}\eta_{ig} + \epsilon_{ipg} \quad (24)$$

where $p = 1, \dots, P$ and P is the number of observed indicator variables, $g = 1, \dots, G$ and G is the number of groups, $i = 1, \dots, N_g$ where N_g denotes the number of independent observations in group g, η_{ig} denotes a latent variable, and it is assumed that $\epsilon_{ipg}\widetilde{N}(0, \theta_{pg}), \eta_{ig}\widetilde{N}(\alpha_g, \psi_g)$. In this context, a configural model can be estimated where all the intercepts v_{pg} and factor loading parameters λ_{pg} are not constrained however, given that the factor means and variances are not identified in this model, the factors η are not comparable across groups and are likely to be on a different scale in each group making it impossible to compare factor scores across individuals within different groups. The alignment optimization though, can estimate the model stated in the equation because it does not assume measurement invariance and it can also estimate the factor mean and variance parameters within each group while obtaining the most optimal measurement invariance pattern through the implementation of a simplicity function that is similar to the rotation criteria used in exploratory factor analysis (EFA) (Asparouhov & Muthén, 2014; Halamová et al., 2019).

Specifically, the alignment optimization allows for the estimation of all the parameters $v_{pg}, \lambda_{pg}, \alpha_g, \psi_g$ by incorporating the assumption that the number of noninvariant measurement parameters and the amount of measurement noninvariance can be held to a minimum into the estimation. The first step in the alignment approach involves the estimation of the configural model where the group factor mean and factor variance are set to equal 0 and 1, respectively that is, $\alpha_g = 0,\ \psi_g = 1$ for every g while

all loading and intercept parameters are freely estimated. This model is referred to as the base model $M0$ and is the best fitting model among all the multiple-group factor analysis models given that it does not have across-group parameter restrictions. The final aligned model has the same fit as $M0$ because even though the aligned model aims to minimize the amount of noninvariance, it does not compromise model fit. In this sense, the relationship between the $M0$ model and the final aligned model is parallel to the relationship between an unrotated and a rotated model in EFA, which simplifies the loading matrix without compromising model fit (Asparouhov & Muthén, 2014; Byrne & van de Vijver, 2017; Kline, 2016; Marsh et al., 2018).

The estimates of the $M0$ model are denoted by $v_{pg,0}, and\ \lambda_{pg,0}$ and the configural $M0$ model transforms the factor within each group to mean zero and variance 1,

$$\eta_{g0} = {(\eta_g - \alpha_g)}\Big/{\sqrt{\psi_g}} \tag{25}$$

the variance and mean indicators can be re-expressed as:

$$V(y_{pg}) = \lambda_{pg}^2 \psi_g = \lambda_{pg,0}^2 \tag{26}$$

$$E(y_{pg}) = v_{pg} + \lambda_{pg}\alpha_g = v_{pg,0} \tag{27}$$

$$\lambda_{pg,0} = \lambda_{pg}\sqrt{\psi_g} \tag{28}$$

$$v_{pg,0} = v_{pg} + \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}\alpha_g \tag{29}$$

For every set of parameters α_g and ψ_g, there are intercept and loading parameters v_{pg} and λ_{pg} that result in the same likelihood as the configural model. These parameters are obtained by:

$$\lambda_{pg,1} = \frac{\lambda_{pg,0}}{\sqrt{\psi_g}}, \tag{30}$$

$$v_{pg,1} = v_{pg,0} - \alpha_g \frac{\lambda_{pg,0}}{\sqrt{\psi_g}} \quad (31)$$

and the goal is to choose α_g and ψ_g to minimize the amount of measurement noninvariance that is, the total loss/simplicity function F that accumulates the total measurement noninvariance is minimized with respect to α_g and ψ_g:

$$F = \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(\lambda_{pg_1,1} - \lambda_{pg_2,1}) + \sum_p \sum_{g_1<g_2} w_{g_1,g_2} f(v_{pg_1,1} - v_{pg_2,1}) \quad (32)$$

For every pair of groups and every intercept and loading parameter, the difference between the parameters scaled through the component loss function (CLF) f is added to the total loss function. The CLF can be given by

$$f(x) = \sqrt{\sqrt{x^2 + \epsilon}} \quad (33)$$

where ϵ denotes a small number such as 0.01. The function is approximately equal to $\sqrt{|x|}$ and this leads to no loss if $x = 0$. However, if $x < 1$ the loss is amplified and if $x > 1$, the loss is attenuated therefore, the total loss function F is minimized at a solution where there are a few large noninvariant measurement parameters and many approximately invariant measurement parameters instead of many medium-sized noninvariant measurement parameters. Moreover, the weight factor w_{g_1,g_2} in F is set to reflect the group size and the amount of certainty in the group estimates for a specific group:

$$w_{g_1,g_2} = \sqrt{N_{g_1} N_{g_2}} \quad (34)$$

With the implementation of this weight factor, larger groups contribute more to the total loss function than smaller groups (Asparouhov & Muthén, 2014).

To summarize, the alignment optimization aims to minimize the amount of measurement noninvariance by estimating factor means (α) -that are allowed to vary

across groups- and factor variances (ψ) while imposing restrictions to optimize a simplicity function F at a few large noninvariant parameters and many approximately invariant parameters. In the alignment optimization of the simplicity function, the factor means α_j and variances ψ_j are free parameters and the same fit as the configural model is obtained for every set of factor means and variances where the factor loadings λ_j and intercepts ν_j are defined as:

$$\lambda_j = \frac{\lambda_{j,configural}}{\sqrt{\psi_j}}, \tag{35}$$

$$\nu_j = \nu_{j,configural} - \frac{\alpha_j \lambda_{j,configural}}{\sqrt{\psi_j}} \tag{36}$$

This method assumes that most of the parameters are invariant and only a minority are non-invariant therefore, if this assumption is not met, biased estimations are likely to occur (Kim et al., 2017; Lamm et al., 2019; Lomazzi, 2018; Muthén & Asparouhov, 2013; Muthén & Asparouhov, 2014; Muthén & Asparouhov, 2018).

In terms of the algorithm that is used to determine invariance, the procedure begins by finding the largest invariant set of groups for each measurement parameter so that in every group within that invariant set of groups, the measurement parameter is not statistically significant from the average value for the parameter across all the groups in that invariant set whereas, for each group that is not in the invariant set, the parameter is statistically significantly different from the average. The algorithm involves multiple pairwise comparisons, but first, it determines a starting set of invariant groups and performs a pairwise test for each pair of groups connecting two groups if the p value is larger than .01. Then, the largest connected set for that parameter is selected to be the starting set of groups. The average parameter is computed using the current invariant set

and then, a test of significance is conducted for each group to compare the parameter value for each group with the current average. If the p value is above .001, the group is added to the invariant set otherwise, the group is removed from the invariance set; the process is repeated until the invariant set stabilizes (Asparouhov & Muthén, 2014; Byrne & van de Vijver, 2017).

Recently, some studies have been implemented to evaluate the performance of the alignment optimization when applied to large-scale educational data. For instance, Byrne and van de Vijver conducted a study on 2017 to evaluate the performance of the alignment procedure when applied to test for measurement invariance on data from a two-factor Family Values Scale across 27 countries. The authors identified five major advantages of this procedure when applied to a large number of groups: (a) it enables tests for measurement invariance and latent mean differences when using large-scale data, (b) it allows for the estimation and comparisons of latent means even when the measures are not fully invariant, (c) it simplifies the comparative analyses, (d) it makes possible to conduct tests for invariance in sub-populations within countries as in the case of cross-cultural research, and (e) it results in refined scales and unbiased statistical estimations of groups means while adjusting for sampling errors and missing data.

Despite the advantages that have been reported for the alignment optimization, Marsh et al. (2018) pointed some limitations such as: the method can only be applied to test a limited number of CFA models, it cannot incorporate cross-loadings, covariates, or tests of SEM, and it can only be used as a mere exploratory tool. Similarly, Muthén and Asparouhov (2014) pointed out the alignment method can lead to parameter biases as (a)

the degree of measurement non-invariance increases, (b) the sample size decreases, and (c) the number of groups increases to more than 60.

In an effort to overcome these limitations Marsh et al. (2018) introduced the alignment-within-CFA (AwC) approach that transforms the traditional alignment method from an exploratory to a confirmatory tool allowing researchers to perform analysis similar to exploratory structural equation modeling (ESEM) within CFA. The steps to conduct AwC include: (a) test a standard multiple-group factor analysis alignment, and (b) reconfigure it as a standard CFA model using the final estimates from the alignment solution as starting values with appropriate fixed and free parameter estimates so that the AwC solution can be equivalent to the multiple-group factor analysis alignment with respect to the number of estimated parameters, goodness of fit, and factor structure.

Regarding model identification, typically one item from each factor is randomly selected to be the reference indicator so that its factor loading, and intercept are fixed to the estimated values from the alignment solution (the starting values are provided by Mplus). The alignment solution as well as the AwC solution, have the same degrees of freedom, the same Chi-square and goodness of fit statistics as the configural MG-CFA model (Marsh et al., 2018).

Marsh et al. conducted a study on 2018 where they applied the traditional multiple-group factor analysis alignment models to evaluate cross-cultural differences in the latent means of the scales related to motivational and engagement constructs in science from PISA 2006 as well as the relationships among the motivational factors and three covariates: gender, science achievement, and socioeconomic status (SES). The authors applied AwC to multiple indicators multiple cause (MIMIC) and MG-CFA

models and used the robust maximum likelihood estimator (MLR). Moreover, they applied corrected standard errors and model fit statistics to control for the nesting of students within schools (using the TYPE=COMPLEX option in Mplus) and used the default option in Mplus for the traditional multiple-group factor analysis alignment model where the latent factor mean and variance of one group are fixed to be 0 and 1, respectively. The authors ran all the analyses involving achievement measures separately for each of the five plausible values and used the full information maximum likelihood (FIML) on each of the five data sets (each one based on different plausible values) to handle the missing data from remaining items. Then, they obtained final parameter estimates, standard errors, and goodness of fit statistics through the automatic aggregation procedure performed in Mplus for multiple imputation.

The analyses conducted by Marsh et al. (2018) on the PISA data included several steps: (a) preliminary CFA to evaluate the factor structure and the correlations with the covariates on the total group, (b) traditional CFA test of measurement invariance of factor structure across countries where increasingly restricted tests of measurement invariance were conducted across 30 countries starting with the configural invariance model that served as the baseline model and continuing with the metric invariance model to conclude with the metric invariance model, (c) implementation of the MG-CFA model with the alignment method given the lack of evidence in the previous step to support scalar invariance; the goal in this step is to perform analysis that allow for the comparison of latent means across countries, (d) implementation of AwC to test invariance constraints on combinations of uniqueness, factor variance and factor covariances, and (e) integration of multiple-group and MIMIC approaches to evaluate the variation of the

relationship between the covariates and each of the motivational constructs across countries by regressing the constructs on each of the three covariates and evaluating the differences across the 30 countries in the context of an SEM analysis. The authors concluded that alignment augmented by AwC provides applied researchers from diverse disciplines considerable flexibility to address substantively important issues when the traditional CFA approach to measurement invariance scalar model does not fit the data.

In summary, the alignment optimization has recently been proposed as a suitable alternative to test for measurement invariance that can handle the complex features of data from ILSAs and several advantages have been associated to the method. However, the method has some limitations including that when the sample sizes are small and the level of noninvariance is large across the groups, the parameters could be biased. Therefore, a modification to the method was introduced on 2018 to overcome these limitations known as the alignment within CFA that according to the empirical evidence, seems to be a proper alternative to handle the limitations of the traditional approach. However, more sound empirical evidence should be collected to evaluate the performance of this new approach.

2.4.2.4 Multilevel Factor Mixture Modeling. This modeling technique involves latent classes and factors. Latent classes emerge when there are unknown heterogeneous subpopulations so that when there is measurement noninvariance across a large number of groups, the groups with the same measurement model or parameters will cluster together forming the latent classes that are formed out of groups (e.g., cultures, countries) and not out of individuals within groups which is why latent classes should be specified

for groups at the between level. Since a measurement model is built for each latent class, models are specified as:

$$[Y_{ij}|c_j] = \tau_{jc} + \Lambda_{jc}\eta_{ijc} + \varepsilon_{ijc} \quad (37)$$

where Y_{ij} = observed score of individual i within group j, cj = latent class where group j belongs so that given the latent class, the relation of the observed variables Y with latent factors η is modeled, τ_{jc} = intercept for group j in latent class c, Λ_{jc} = factor loadings for group j in latent class c, η_{ijc} = common factor scores for individual i within group j in latent class c and ε_{ijc} = residuals for individual i within group j in latent class c . A multinomial regression model is used to estimate the latent class membership of group j and the log odds of being a member of class C over a reference class 1 is modeled with between-level observed predictors X_j so that

$$\log\left[\frac{P(c_j=C|X_j)}{P(c_j=1|X_j)}\right] = v_c + \Gamma_C X_j \quad (38)$$

where $v_C \; and \; \Gamma_C$ = intercepts and regression coefficients for latent class C, respectively that take a value of zero for the reference class for identification.

When testing for measurement invariance, latent classes are treated as between-level latent categorical variables and since the number of latent classes and the location of heterogeneity are not known, a series of models should be constructed and the level of measurement invariance should be determined by comparing the series of models (e.g., one class model, two-class configural, two-class metric, two-class scalar). In this scenario, there is only one CFA model that is specified at the within-level for all groups and the model fit for all the models are compared simultaneously and the best fitting model is selected (Kim et al., 2017).

Multilevel factor mixture modeling informs how the groups cluster together and identifies class membership for each group moreover, it allows for the assessment of sources of noninvariance through the modeling of potential observed covariates and class membership. This approach also performs well under large number of groups however, several comparisons must be performed, and different sets of starting values and larger number of iterations are often required because nonconvergence is very common. Another potential issue with the implementation of this approach is that the stability of class membership cannot be guaranteed across the models so that the members that belong to each class can be different from one model to another (Kim et al., 2017). Given these limitations, the method has rarely been implemented in the context of the evaluation of measurement invariance applied to data from ILSAs and given the lack of stability in terms of class membership its use is not recommended.

2.4.2.5 Exploratory Structural Equation Modeling (ESEM). Exploratory structural equation modeling (ESEM) has been suggested as an alternative to multi-group CFA that combines the strengths of both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) with a standard structural equation modeling approach allowing for a less restrictive testing for the equivalence of factorial structures where all non-target factor loadings and error covariances are freely estimated (Byrne & van de Vijver, 2017; Fischer & Karl, 2019).

ESEM differs from CFA in that all factor loadings are estimated subject only to identification constraints thus, items are free to cross-load on more than one factor (Marsh et al., 2011). The implementation of ESEM typically begins with an EFA to formulate the factor structure, then an ad hoc procedure is used to mirror the EFA

structure as an SEM model including a CFA measurement specification that will be further tested (Byrne & van de Vijver, 2017; Fischer & Karl, 2019). For the purpose of testing for measurement invariance, two general steps are conducted. First, an unconstrained factor structure is estimated, and then the resulting structure is rotated usually through oblique or orthogonal rotations that allow for the specification of a-priori assumptions on the factor structure (Scherer et al., 2016).

One of the advantages of ESEM is that it has been proposed as a suitable method to evaluate complex non-cognitive constructs such as bullying where it is very likely that the relations among factors are inflated (Marsh et al., 2011). For instance, Marsh et al. conducted a study on 2011 to better understand bullying and victimization in high schools. They tested the psychometric properties and measurement invariance of an instrument that measured multiple bully and victim factors. They explored the utility of ESEM in the assessment of a bullying and victimization instrument that included six scales where students indicated the frequency with which they engaged in a series of behaviors using a 6-point likert scale. They used an ESEM approach with the full-information robust maximum likelihood estimator (MLR) to conduct multiple group analysis where the ESEM solution is estimated separately for each group and some parameters are constrained to invariance across groups. Moreover, they assessed sample size-independent goodness of fit statistics: RMSEA, TLI and CFI, and found evidence of good model fit and when compared to the CFA the fit was better suggesting that the ESEM detected the distinction among the facets of bullying and victimization.

ESEM has also been used to evaluate measurement invariance on some of the non-cognitive measures from PISA. For example, Meng et al. (2018) conducted a study

on measurement invariance of the ICT engagement scale from PISA 2015 comparing China and Germany. They implemented exploratory structural equation modeling (ESEM) which according to the authors is a superior statistical technique when compared to multigroup confirmatory factor analysis for cross-cultural comparisons. They first implemented ESEM to measure the underlying latent construct for each country using the following criteria for goodness of fit: CFI and TLI with values higher than 0.95 along with SRMR and RMSEA with values less than 0.05. Then, they tested for configural, metric and scalar invariance using the following goodness of fit criteria: $\Delta CFI \leq 0.010, \Delta TLI \leq 0.010, \Delta RMSEA \leq 0.015, and\ \Delta SRMR \leq 0.015$. They implemented the robust maximum likelihood estimation method and according to the results, the scalar level of invariance was achieved. However, the authors pointed the need to collect more evidence to determine if the scale has broad applications in different countries with different cultures.

Given the exploratory nature of ESEM, it is not likely to obtain the proper loading pattern when modeling a latent construct featuring many indicators and factors thus, the method could lead to biased estimates and interpretations (Asparouhov & Muthén, 2009). In this sense, the approach is not recommended for the evaluation of measurement invariance in the context of ILSAs.

2.4.2.6 Bayesian Approximate Testing for Measurement Invariance. Bayesian statistics in general express the uncertainty about a population value of a particular parameter through a probability distribution of possible values that is known as the prior distribution that is specified independently from the data. After data collection, the prior distribution is combined with the likelihood of the data to generate a posterior

distribution that describes the uncertainty about the population values and the variance of the posterior distribution is usually smaller than the variance of the prior distribution since the observed data reduces the uncertainty about possible population values (Hox et al., 2012).

In this context, the approximate Bayesian measurement invariance has been suggested as an alternate approach to the traditional evaluation of exact measurement invariance that relaxes the assumptions of full measurement invariance allowing for small variation of factor loadings and intercepts across groups. In the cases where several groups are compared, strict assumptions from full measurement invariances lead to poor fit. The Bayesian approach can be used to allow for a small variation where it is assumed that the parameters are random, and the uncertainty is incorporated into parameter estimation using resampling techniques. Prior knowledge on parameters can be incorporated specifying a prior distribution of a parameter in the model thus, when testing for measurement invariance a plausible range of differences in factor loadings and intercepts between groups can be specified in advance and the posterior distribution of a factor loading is estimated as a function of the prior distribution based on the likelihood function from the data (Cieciuch et al., 2014). The main difference between the approximate and traditional measurement invariance is that in the latter some parameters are constrained to be exactly equal, and others are released completely while in the former even though all parameters are constrained, their restrictions are not strict allowing for approximate equality (Cieciuch et al., 2014).

In terms of the implementation, measurement invariance is tested through four steps: (a) a multiple group CFA model is tested without equality constraints on factor

loadings or intercepts (configural model) using the Bayesian estimation method where the factor mean and variance of all groups are set to 0 and 1, respectively for identification purposes while all factor loadings and intercepts are freely estimated; (b) prior variance that is approximately invariant for the differences in factor loadings and intercepts is determined; (c) a series of approximate metric measurement invariance models are built including the prior variance, and the best fitting model is chosen from model comparisons so that approximate metric invariance holds if the selected model is smaller than or equal to the predetermined value; moreover intercepts are freely estimated for the approximate metric invariance models and factor means are set to zero for identification purposes; and (d) if approximate metric invariance holds, the Bayesian approximate measurement invariance (scalar model) is tested for full approximate measurement invariance by repeating the previous step for the intercept differences so that if the model with the prior variance smaller than or equal to the predetermined prior variance for intercept differences is selected, then measurement invariance holds. The factor mean for one group is set to zero while the remaining factor means, and variances are freely estimated for identification purposes and the models are evaluated using model evaluation strategies for Bayesian analysis like the posterior predictive checking (Kim et al., 2017).

In terms of the performance of this approach, empirical studies have provided mixed results. For instance, Cieciuch et al. conducted a study on 2014 with the aim to assess the measurement invariance of a scale to measure human values using an approximate (Bayesian) approach for testing measurement invariance. The authors used mixture modeling where besides the latent variables there are also latent categorical

variables that describe membership of test takers to a specific class which represent homogeneous subpopulations from the target heterogeneous populations. In their study, the subpopulations were given by countries (eight countries in total) and thus, they evaluated a single class mixture model. Evaluation of model fit identifies if actual deviations are larger than the deviations the researcher allows in the prior distribution through the posterior predictive probability value and the confidence interval for the difference between observed and replicated Chi-square values therefore, Bayesian models fit the data when the posterior predictive probability value is higher than zero and the confidence intervals contain zero. The authors compared their results to results from an exact approach to measurement invariance and found that the less restrictive method (approximate Bayesian approach) resulted in stronger invariance.

On the other hand, several limitations have been associated to this approach. For example, a simulation study by Hox et al. (2012) collected evidence suggesting that the estimations under this approach can be inaccurate when the sample sizes are small.

Other limitations are related to the fact that the approach usually requires the construction of several models with different levels of prior variances to identify the best fitting model. Moreover, the model evaluation criteria are not well established, and the execution time increases as sample size becomes larger (Kim et al., 2017). Another salient limitation is that the prior for the variance parameter entails an assumption of measurement non-invariance that if not met, increases the likelihood that the target latent construct is estimated using potentially biased item difficulty and population parameter estimates (van de Vijver et al., 2019).

Regarding the suitability of the approach to handle data from ILSAs, it has been reported that the evaluation of measurement invariance becomes cumbersome when analyzing large-scale data involving many groups and it is usually the case that full measurement invariance does not hold in those scenarios (van de Vijver et al., 2019). In summary, empirical evidence in general, suggests that the approach might not be suitable to handle the features of data from PISA.

To conclude, there are several statistical approaches that have been proposed to test for measurement invariance. In general, the approaches have been classified into two categories: traditional and modern. Modern approaches are suitable alternatives that aim to overcome the limitations of traditional approaches which despite their well-known disadvantages are still being used for the evaluation of measurement invariance.

Among the modern approaches, the multilevel structural equation modeling (MSEM) and the alignment optimization method have many advantages that in general make them a suitable alternative to handle the complex features of data from ILSAs. In the case of the alignment optimization method, it has not been widely implemented in the context of educational international assessments and the available empirical evidence positions the method as a promising alternative that could help overcome the limitations that are usually found when assessing measurement invariance in ILSAs.

Theoretical and empirical evidence seem to point that the lack of measurement invariance that is typically reported in studies analyzing data from PISA, could be due to the deficiencies in the statistical approaches that have been implemented and that are not properly handling the features of the data. Therefore, this dissertation will provide sound empirical evidence about the performance of the alignment optimization method and the

MSEM approaches to evaluate measurement invariance in cognitive and non-cognitive measures from PISA 2018.

CHAPTER

III METHOD

This dissertation involved secondary data analyses that were performed on the Programme for International Student Assessment (PISA) student-questionnaire data from 2018 with the aim to offer evidence about the extent to which PISA provides invariant measures of reading literacy and two non-cognitive measures for immigrant students from diverse cultural and linguistic backgrounds across the countries that tested more than 300 immigrant students.

3.1 Sample

Secondary data from 218,315 students were analyzed. Of them, 169,651 were native and 48,664 were immigrants. These data (see Table 1) were retrieved from the OECD-PISA website (https://www.oecd.org/pisa/data/2018database/).

The distribution of students across countries selected for this study and immigration status is shown in Table 1.

Table 1

Distribution of Students per Country and Immigration Status

Country	Immigration Status		Total
	Native	Immigrant	
Germany	1736	422	2158
Switzerland	2189	1104	3293
Netherlands	3092	389	3481
Macao	1358	2287	3645
France	3621	524	4145
Luxembourg	1967	2217	4184
Ireland	3504	706	4210
United States	3388	890	4278
Serbia	3888	429	4317
Brunei Darussalam	4024	419	4443
Sweden	3711	761	4472
Estonia	4119	462	4581
Slovenia	4314	362	4676
New Zealand	3504	1194	4698
Austria	4020	976	4996
Norway	4417	593	5010
Croatia	4596	468	5064
Greece	4563	526	5089
Hong Kong	3360	2041	5401
Denmark	4683	1070	5753
Costa Rica	5220	576	5796
Singapore	4757	1418	6175
Belgium	5695	1089	6784
Italy	7330	720	8050
Australia	7297	2700	9997
Qatar	4071	6131	10202
United Kingdom	9775	1428	11203
Kazakhstan	13257	1013	14270
United Arab Emirates	6939	8404	15343
Canada	13131	4573	17704
Spain	22125	2772	24897
Total	169651	48664	218315

3.1.1 Inclusion Criteria

The countries to be analyzed were selected based on the following criteria:

- Countries that host large population of immigrants according to the United Nations (2017).
- Countries that reported PISA data from at least 300 immigrant students or more to meet the requirements for the proper implementation of the statistical modeling techniques.

Given these criteria, 31 countries were included in the analyses: Australia, Austria, Belgium, Brunei Darussalam, Canada, Costa Rica, Croatia, Denmark, Estonia, France, Germany, Greece, Hong Kong, Ireland, Italy, Kazakhstan, Luxembourg, Macao, Netherlands, New Zealand, Norway, Qatar, Serbia, Singapore, Slovenia, Spain, Sweden, Switzerland, United Arab Emirates, United Kingdom, and United States.

3.1.2 Exclusion Criteria

Exclusion criteria included:

- Countries with less than 300 immigrants.
- Countries with missing data in the reading subscales.
- Cases with no report of immigration status.
- Cases with missing data in any items from the non-cognitive scales.

The average age of the students in the sample was 15.25 (*SD*= .29) and the sex and grade at the time of the assessment are described in Table 2.

Table 2

Sociodemographic Characteristics of the Students

	Native		Immigrant	
	n	%	*n*	%
Sex				
Female	86,628	51.1	24,623	50.6
Male	83,023	48.9	24,041	49.4
Grade				
Grade 7	303	0.2	227	0.5
Grade 8	4,160	2.5	2,114	4.3
Grade 9	43,489	25.6	12,078	24.8
Grade 10	98,496	58.1	25,828	53.1
Grade 11	17,385	10.2	6,955	14.3
Grade 12	1,791	1.1	484	1.0
Grade 13	7	0	1	0

3.1.3 Sampling Procedures

The target population in PISA includes 15-year-old full or part-time students attending educational institutions and enrolled in grades seven or higher. The sampling design implemented in all the countries (except Russia) was a two-stage stratified sample design where the first-stage sampling units were the schools that had 15-year-old students; schools were systematically sampled from a national list of PISA-eligible schools with probabilities proportional to a measure of size which is a function of the estimated number of eligible 15-year-old students enrolled in each school.

This sampling strategy is known as systematic probability proportional to size sampling. The eligible schools were first assigned to mutually exclusive groups based on school characteristics or explicit strata to improve the accuracy of the sample-based estimated (OECD, 2018a). The second-stage sampling units consisted of the students within the sampled schools who were selected with equal probability to be part of a target cluster of either 42 students in countries that participated in the computer-based

assessment or 35 students in countries participating in the paper-based assessment (OECD, 2018a).

3.2 Instrument

3.2.1 Programme for International Student Assessment (PISA)

PISA was developed by the Paris-based Organization for Economic Co-operation and Development (OECD) with the aim to provide international comparative educational data suitable to be used for policy-making purposes (He et al., 2019; Meng et al., 2018; OECD, 2016; Rubinstein-Avila, 2016; Volante et al., 2017). PISA is administered on a three-year basis to 79 participating countries throughout the world and focuses on core school subjects of reading, mathematics, and science. The goal is to determine how well students can extrapolate what they have learned and apply that knowledge in unfamiliar settings (OECD, 2019a).

PISA also collects information about the students' home background that involves non-cognitive variables, and in combination with the major domains, PISA provides three outcomes:

- Indicators that provide a profile of the knowledge and skills of the students.
- Indicators that show the relationship between the skills and demographic, social, economic, and educational variables.
- Indicators that show changes in the relationships between student, school and system-level background variables and the outcomes (OECD, 2019a).

The assessment in 2018 was mainly delivered through computer-based format and the total time of assessment per student was 2 hours. In terms of the item format, test

items include multiple-choice questions and constructed-response questions where students are required to construct their answers (OECD, 2019a).

3.2.1.1 Cognitive Measures. As previously mentioned, PISA evaluates three major cognitive domains: reading, mathematics and scientific literacy. Given that the focus of this dissertation will be on the measure of reading literacy, details will be provided for that domain area.

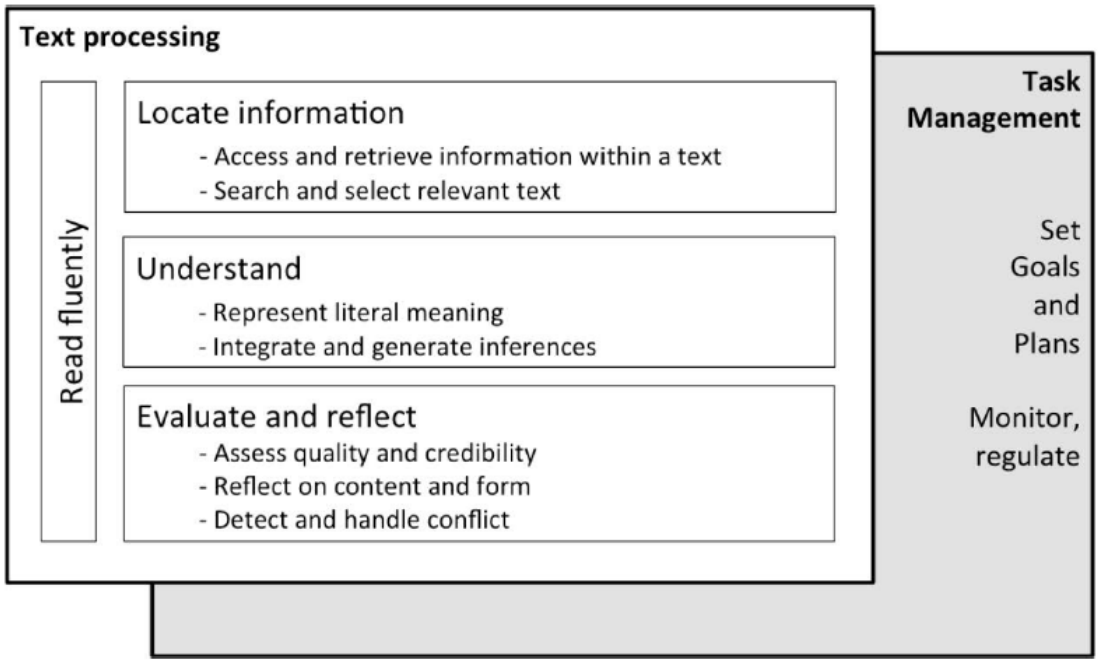
3.2.1.1.1 Reading Literacy. The major domain in PISA 2018 was reading literacy defined as "understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society" (OECD, 2019a, p. 28).

The reading literacy domain was developed according to: (a) reader factors that include motivation, and prior knowledge, among others; (b) text factors, that refer to the format of the text, the complexity of the used language, and the number of pieces of text; and (c) task factors that include the potential time, the goals of the task, and the complexity of the reading task (OECD, 2019a).

The major aim of the reading literacy assessment was to measure the students' mastery of reading processes and two broad categories of reading processes were defined for PISA 2018: text processing and task management as seen in Figure 1.

Figure 1

PISA 2018 Reading Framework Processes



Note. From OECD (2019a). *PISA 2018 Assessment and Analytical Framework*. PISA, OECD Publishing. https://doi.org/10.1787/b25efab8-en. Copyright 2019 by the Organisation for Economic Co-operation and Development.

Reading fluency refers to the ability to read words and texts with precision and in an automatic way and to process the words and texts to comprehend the meaning of a text. Reading fluency involves three major processes:

1. Locating information. Relates to the ability to carefully read an entire text to comprehend main ideas and reflect on the text as a whole. The ability to locate information is a mandatory component of reading especially when using complex digital information (i.e., websites) that involves the ability to access and retrieve information from a text (i.e., locate information from tables), regulate the reading speed and depth of processing, and search and select relevant information.

2. Understanding. Refers to the ability to make a mental representation of the information in the text and the integration of the contents provided in the text with the readers' prior knowledge through inference processes.
3. Evaluating and reflecting. Refers to the ability to reflect on the content of the text and critically assess its quality and the validity of the information.

PISA 2018 selected these three processes as the reporting subscales (OECD, 2019a).

On the other hand, the task management processes involve the ability to represent the reading demands of a situation, set up reading goals, monitor progress, and self-regulation towards the reading goals (OECD, 2019a).

Regarding the reading literacy texts, PISA 2018 classified the texts into four categories as shown in Table 3. Finally, regarding the response modes, PISA 2018 included five response modes: (a) click on a choice such as single-selection multiple choice, multiple-selection multiple choice, complex multiple choice or click on an image, (b) numeric entry, (c) text entry, (d) select from a drop-down menu, and (e) drag and drop (OECD, 2018a).

Table 3

PISA 2018 Reading Literacy Texts

Categories	Examples
Source	• Single unit of text. • Multiple units of texts where each has a different author.
Organization and navigation	• Static texts with simple organization and low density of navigation tools. • Dynamic texts with complex organization and higher density of navigation tools.
Format	• Continuous texts with sentences organized into paragraphs. • Non-continuous texts composed of several lists or elements. • Mixed texts containing continuous and non-continuous elements.
Type	• Description texts. • Narration texts with information related to objects in time. • Exposition texts with explanations of how different elements relate in a meaningful way. • Argument texts that provide a relationship among concepts. • Instruction texts that provide steps or instructions on what to do. • Transaction texts including letters, emails, or text messages.

Note. From Organisation for Economic Co-operation and Development [OECD]. (2018a). *PISA 2018 Technical Report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/#d.en.423800

Results were described according to the reading proficiency levels reported by PISA for Reading Literacy as shown in Table 4.

Table 4

Score Categories

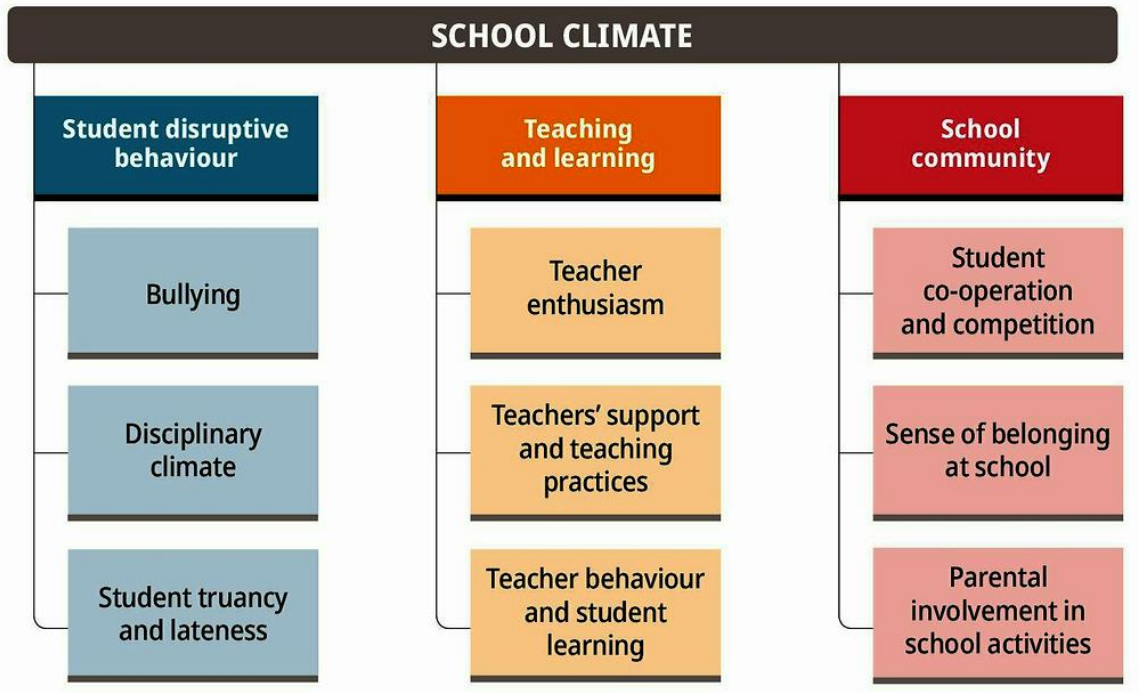
Lower Score Limit	Level	Characteristics of readers
698	6	Make multiple inferences, comparisons and contrasts. Demonstrate full understanding of one or more texts. Integrate information from more than one text. Generate abstract categories for interpretations. Critically evaluate complex texts. Precision of analysis and attention to detail that is inconspicuous in the texts.
626	5	Locate and organize several pieces of information and determining which information in the text is relevant. Critical evaluation, hypothesis-making drawing on specialized knowledge. Understanding of texts whose content or form is unfamiliar. Deal with concept that are contrary to expectations.
553	4	Locate and organize pieces of embedded information. Interpret nuances of language. Understanding and application of categories in an unfamiliar context. Use formal or public knowledge to evaluate a text. Demonstrate accurate understanding of long and complex texts that have unfamiliar content of form.
480	3	Locate and recognize the relationship between several pieces of information that must meet multiple conditions. Integrate several parts of the text to identify the main idea, understand a relationship, or interpret the meaning of a word or phrase. Connections, comparisons, and explanations are needed at this level. Evaluate features of a text. Demonstrate understanding of the text with respect to familiar knowledge.
407	2	Locate one or more pieces of information that need to be inferred. Recognize main idea of the text, understand relationships or interpret meaning within a limited part of the text when the information is not prominent. Make low-level inferences. Make comparisons or contrasts based on a single feature in the text. Make connections between the text and outside knowledge by drawing on personal experience.
335	1a	Locate independent pieces of explicitly stated information. Recognize main theme or author's purpose in a text. Make simple connections between information in the text and common knowledge.
262	1b	Locate a single piece of explicitly stated information in a prominent position in a short syntactically simple text with a familiar context. Interpret texts by making simple connections between adjacent pieces of information.

Note. From Organisation for Economic Co-operation and Development [OECD]. (2019a). *PISA 2018 Assessment and Analytical Framework*. https://www.oecd-ilibrary.org/sites/5c07e4f1-en/index.html?itemId=/content/component/5c07e4f1-en

3.2.1.2 Non-cognitive Measures. Besides the three core domain areas, PISA also collects contextual information through a student questionnaire. PISA 2018 offered five additional questionnaires: computer familiarity questionnaire, well-being questionnaire, educational career questionnaire, parent questionnaire, and teacher questionnaire. Within the well-being questionnaire, PISA includes three measures of school climate with sub-dimensions as shown in Figure 2.

Figure 2

School Climate Questionnaires PISA 2018



Note. From OECD (2019b). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*, PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en. Copyright 2019 by the Organisation for Economic Co-operation and Development.

Two measures from the school climate measure were selected to be analyzed: bullying and sense of belonging at school, given their relevance and potential impact on the educational experiences of first-generation immigrant students.

3.2.1.2.1. Sense of Belonging at School. In general, the sense of belonging is related to a natural tendency to maintain interpersonal relationships that are based on acceptance and support thus, students with a sense of belonging at school feel accepted and connected to their peers and school community (OECD, 2019b).

The sense of belonging at school as measured by PISA indicates the extent to which students feel accepted, respected and supported in their social context at school therefore, students with a high sense of belonging at school typically have high motivation, self-esteem and academic achievement (OECD, 2019b).

In PISA, students are asked to indicate how much they agree (strongly disagree, disagree, agree, strongly agree) with each of the following statements about their schools:

- I feel like an outsider (or left out of things) at school.
- I make friends easily at school.
- I feel like I belong at school.
- I feel awkward and out of place in my school.
- Other students seem to like me.
- I feel lonely at school.

(OECD, 2019b).

Based on the students' responses, PISA creates an index of sense of belonging at school that has an average of 0 and standard deviation of 1 (OECD, 2019b).

3.2.1.2.2. Bullying. Bullying is included in PISA as an indicator of the students' connections at school and it is defined as "negative physical or verbal actions that have hostile intent, cause distress to victims, are repeated and involve a power differential between perpetrators and victims" (OECD, 2019b, p. 273).

The measure of bullying contains quantifiable behaviors that are indicators of negative or dysfunctional social relationships and measure three types of bullying: physical, relational and verbal. Students are asked to indicate how often (never or almost never, a few times a year, a few times a month, once a week or more) they have had the following experiences during the 12 months prior to the test:

- Other students left me out of things on purpose (relational bullying)
- Other students made fun of me (verbal bullying)
- I was threatened by other students (verbal/physical bullying)
- Other students took away or destroyed things that belong to me (physical bullying)
- I got hit or pushed around by other students (physical bullying)
- Other students spread nasty rumors about me (relational bullying)

As with the previous measure, PISA provides a general index of bullying based on the answers to these statements. The index has a mean of 0 and standard deviation of 1 and positive values suggest that the student is more exposed to bullying at school than the average student in OECD countries whereas negative values suggest that the student is less exposed (OECD, 2019b, p. 273).

Regarding the dimensionality of the bullying construct, it was assumed that an observed variable x (one of the six items) was the result of a latent response variable

x^*which in this case corresponded to the student exposure to bullying. Therefore, the observed categories of x for each student i corresponded to a specific threshold in the continuum of the latent variable x^* so that:

$x_i =$ "never or almost never" (category 1) if $x_i^* \leq \tau_{i,1}$;

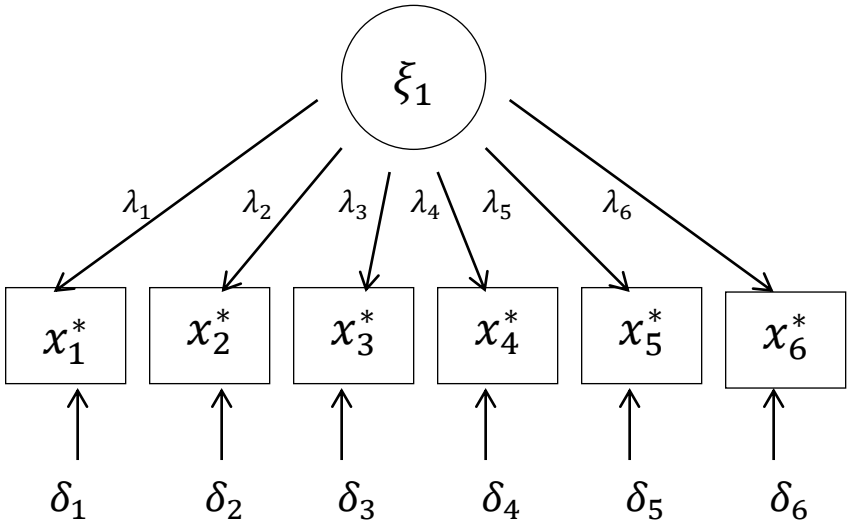
$x_i =$ "a few times a year" (category 2) if $\tau_{i,1} < x_i^* \leq \tau_{i,2}$;

$x_i =$ "a few times a month or once a week or more" (category 3) if $x_i^* > \tau_{i,2}$.

The bullying scale was originally assessed through a model that accounted for the categorical distribution of the data and included these thresholds as parameters to be estimated. Moreover, the model used a theta parameterization where the first factor loading was fixed to 1, the latent variable mean to 0, and the residual variance to 1 across all groups for identification purposes. The graphical representation of the model is shown in Figure 3

Figure 3

PISA Measurement Model for Bullying



Where for any x^*:

$$x_i^* = v_{x^*} + \lambda_{ij}\xi + \delta_i \quad (39)$$

$x_i = c$ if $\tau_c < x_i^* \leq \tau_{c+1}$ (40)

And

ξ_1: latent variable (exposure to bullying)

x^*: latent response variable

x: observed variable

λ_{ij}: factor regression weights

δ_i: measurement error

v_{ic}: thresholds for categories $c = 0,1,2, \ldots, c-1$

(OECD, 2017; Rosen et al., 2013).

This model was obtained after testing the initial model (that included eight indicators) through an exploratory factor analysis (EFA) where the first two items ("I got called names by others" and "I got picked on by other students") did not load onto a unidimensional construct and did not correlate with the rest of the items. Moreover, the estimated averages for those two items varied across countries suggesting that students from different countries interpreted the item contents in different ways therefore, and given the measurement issues with those two items, they were excluded (OECD, 2017).

3.3 Procedure

Given that data collection in PISA involves balanced incomplete block (BIB) spiraling where the total items are divided into small blocks which in turn, are assigned to distinct booklets so that test takers do not respond to more than a fraction of the total number of items (Kaplan, 1995), there are large amounts of missing data. To address this issue, the analyses from reading literacy were performed at the test-level using the data reported for the reading literacy subscales.

The procedure involved three general phases to address the research questions that were stated for this dissertation and collect the pertinent empirical evidence:

1. Descriptive analyses: the dataset was described in terms of (a) the academic and background variables of the sample, (b) the target cognitive and non-cognitive measures, and (c) the psychometric features of the target measures.
2. Evaluation of measurement invariance: analyses were conducted to determine the extent to which the cognitive and non-cognitive measures were invariant across countries and test takers within countries.
3. Evaluation of the relationship between the test performance on the cognitive measure and the non-cognitive measures in the context of structural equation modeling.

3.3.1 Descriptive Analyses

The analyses began with the description of the sample in terms of the following variables: gender and index of economic, social, and cultural status per country. Then, the cognitive and non-cognitive measures were described per country in terms of:

- Item-level means, standard deviations, and discrimination.
- Descriptive statistics at the test level per measure.

These analyses were conducted in RStudio version 1.3.1093.

3.3.2 Evaluation of Measurement Invariance

Tests of measurement invariance were conducted on each of the measures: the two non-cognitive scales, and the reading literacy scale. The analysis of the reading scale was performed on the average of the plausible values provided by PISA.

The analysis was conducted to determine the extent to which each measure is invariant within countries (across first-generation immigrant students and their native peers) and across the countries under analysis. To do so, the two approaches were implemented: the traditional approach (multiple group confirmatory factor analysis) and an alternate approach that according to the literature and empirical evidence has been found to be suitable to handle data from PISA (alignment optimization).

3.3.2.1 Multiple Group Confirmatory Factor Analysis (MGCFA). MGCFA analyses were conducted for each of the three measures under analysis (reading literacy, sense of belonging at school, and bullying) with the aim to identify the extent to which each of the three latent constructs were invariant across countries.

The analyses were conducted (a) at the item level for the two non-cognitive measures and (b) at the test level for the cognitive measure where the scores from three reading subscales were used; this decision was made given the large rate of missing data for the reading literacy items due to the adaptive multistage test design. Therefore, the analyses involving the reading literacy measure were conducted on the average of each plausible value provided per subscale.

The procedure involved the evaluation of configural, metric, and scalar invariance. The evaluation of configural invariance aimed to determine if the indicators were measuring the same factors across countries and to do so, a configural model was specified without imposing any constraints except for those needed for identification purposes (fixing one item per factor to 1). Then, metric invariance was examined to evaluate if the magnitude of the relationships between the indicators and the factors was similar across countries, that is, if the factor loadings were the same. In this analysis, the

factor loadings were set to be equal across countries and the fit of this model was compared to that of the configural model (Pendergast et al., 2017).

Finally, scalar invariance was evaluated. Specifically, equality constraints were imposed on the item intercepts from the reading literacy measure and on the item thresholds from the non-cognitive measures. In the case of the non-cognitive measures, one threshold on each item and two thresholds for one item on the factor were fixed to 1 for identification purposes,[1] moreover, theta parameterization was implemented for the categorical measures thus, the residual variances for the categorical items were estimated while the factor residual variances were not included (Pendergast et al., 2017).

The overall analysis involved three sets of factor analysis formulas:

- Configural model:

$$y_{ij} = \nu_j + \lambda_j f_{ij} + \epsilon_{ij}, \quad E(f_i) = \alpha_j = 0,\ V(f_i) = \psi_j = 1. \tag{41}$$

- Metric model:

$$y_{ij} = \nu_j + \lambda f_{ij} + \epsilon_{ij},\ E(f_j) = \alpha_j = 0,\ V(f_j) = \psi_j. \tag{42}$$

- Scalar model:

$$y_{ij} = \nu + \lambda f_{ij} + \epsilon_{ij},\ E(f_j) = \alpha_j,\ V(f_j) = \psi_j, \tag{43}$$

where i and j denote individual and group, respectively, ν is the measurement intercept, λ denotes a factor loading, f is a factor with mean α and variance ψ, and ϵ denotes the residual variance with a mean of zero and variance θ which is not correlated with f. It can be seen from the formulas that the configural model includes the subscript j for the factor loadings and intercepts whereas, the metric model does not include the subscript

[1] All the items from each scale were tested as the reference item, one at a time, but no changes in the model fit were observed thus, those results were not reported.

for the factor loadings, and finally, the scalar model does not include the subscript neither for the factor loadings nor for the intercepts. Table 5 summarizes the details about the specification and identification of each model in MGCFA.

Table 5

Summary Model Identification MGCFA

	Categorical Measures			**Continuous Measure**		
	Configural	**Metric**	**Scalar**	**Configural**	**Metric**	**Scalar**
Factor loadings freely estimated across groups.	X			X		
Factor loadings constrained to equality across groups.		X	X		X	X
Thresholds freely estimated across groups.	X					
First threshold of each item held equal across groups.		X				
Thresholds constrained to equality across groups.			X			
Intercepts freely estimated.				X	X	
Intercepts constrained to equality across groups.						X
Residual variances fixed to 1.0	X	X	X			
Residual variances freely estimated.				X	X	X
Factor variance freely estimated.	X	X	X	X	X	X
Metric of the factor set by fixing the factor loading of one indicator to 1.0.	X	X	X	X	X	X

The bullying and sense of belonging at school scales included ordinal indicators where the responses were presented in a 4-point Likert scale so that observed responses were not directly related to the target constructs instead, the items were associated to the constructs through $c - 1$ thresholds where c denotes the number of response categories

($c = 4$). Thus, the models in Figures 4 and 5 show that each item (X) is associated with a latent response variable labeled as X^* and the threshold parameters are labeled as τ. The parameters λ and θ denote the factor loadings and error variances, respectively whereas the group (country) factor mean is represented as κ.

Theta parameterization was assumed for the ordinal indicators where the residual variance of each X^* variable was fixed to 1.0 in the first group selected as the reference group (Australia)[2] while error variances in the remaining groups were freely estimated (Kline, 2016). The measurement models that were evaluated through MGCFA are shown below.

[2] The countries were included as reference one at a time and no changes in model fit were observed therefore, those results were not reported.

Figure 4

Bullying Measurement Model for MGCFA

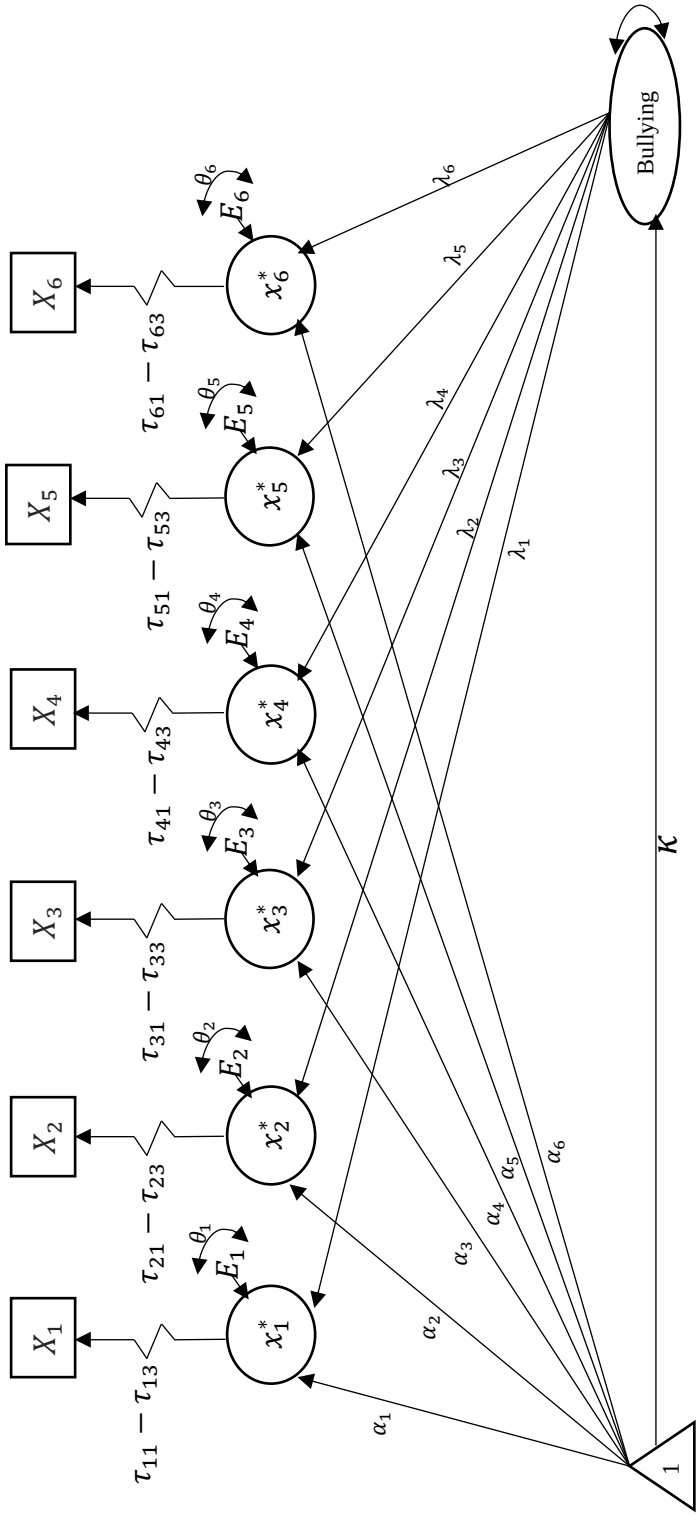


Figure 5

Sense of Belonging at School Measurement Model for MGCFA

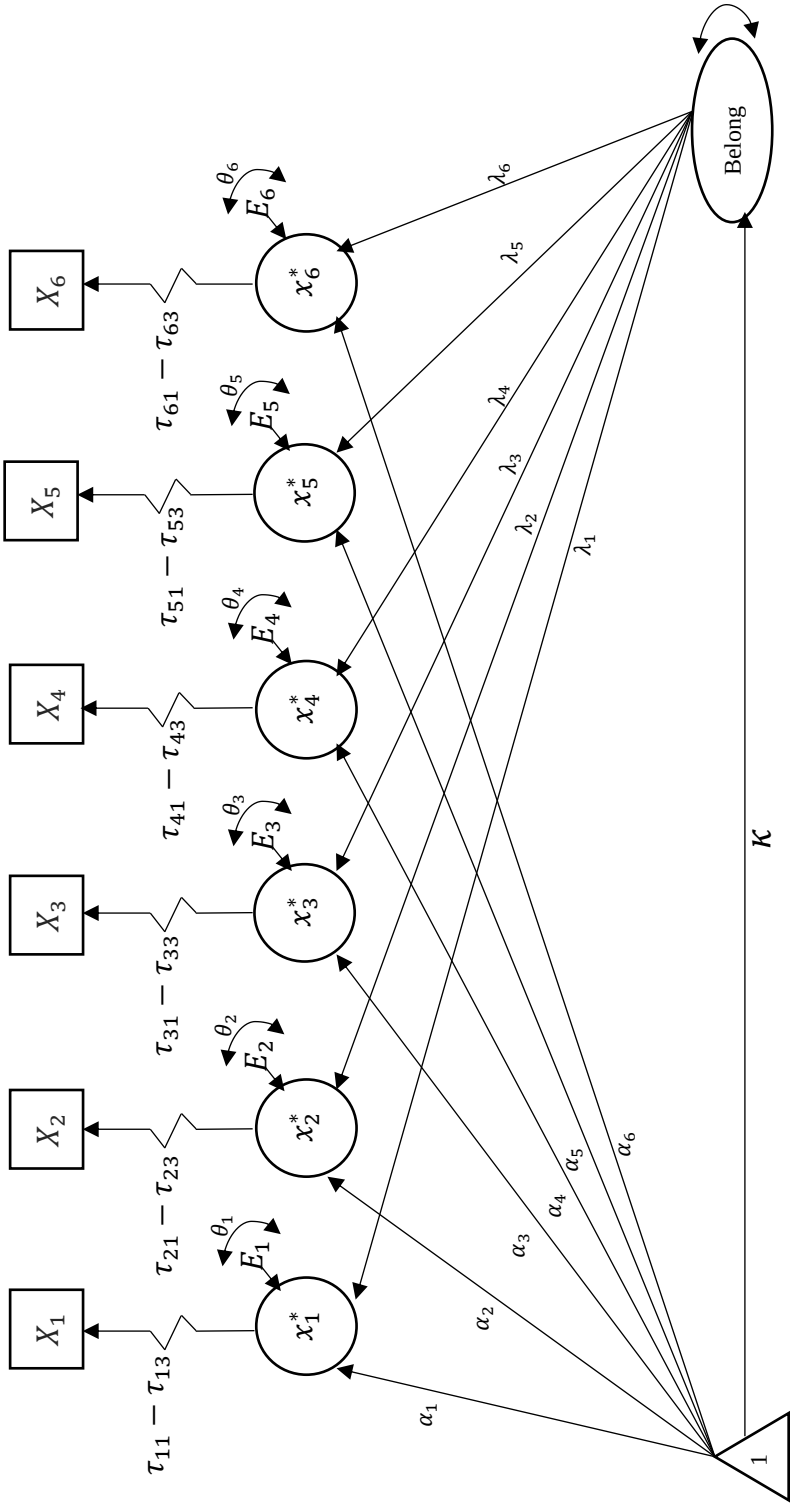
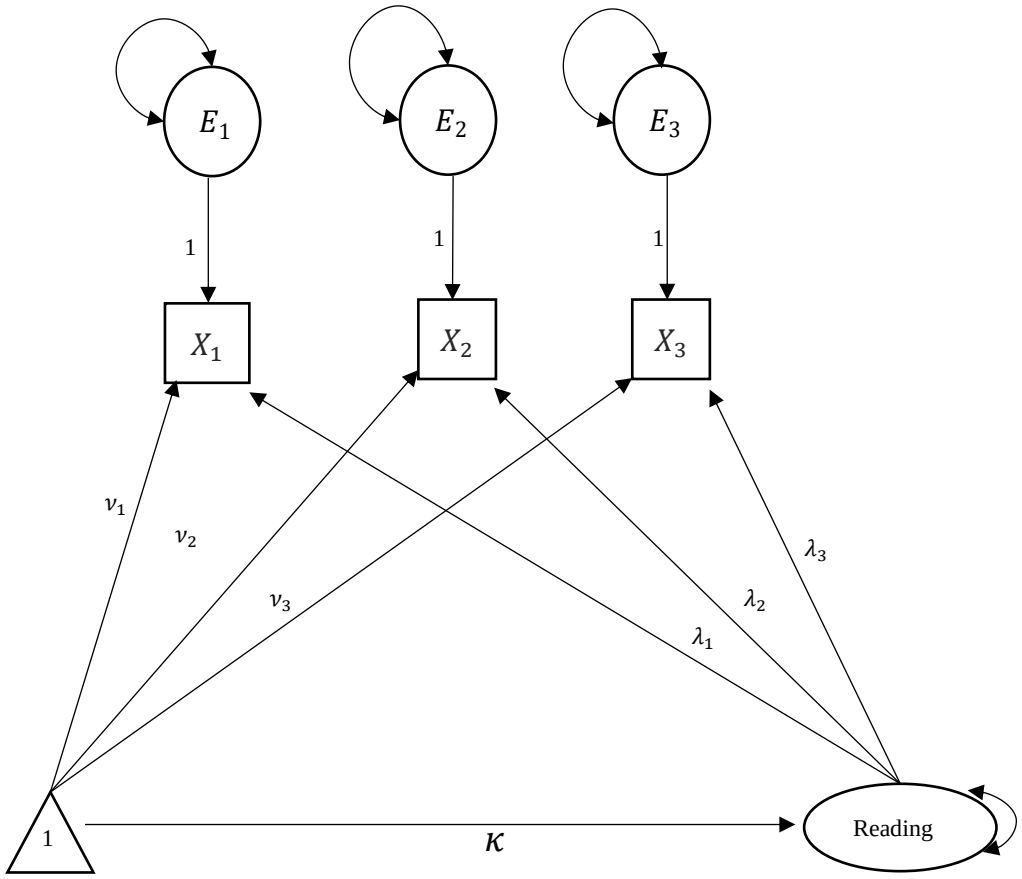


Figure 6

Reading Literacy Measurement Model for MGCFA



The fit of the models was evaluated and compared to one another: configural against metric and metric against scalar. Regarding the estimation method, maximum likelihood (ML) was used for the reading literacy measure and the weighted least squares mean and variance adjusted (WLSMV) was used for the non-cognitive measures.

The goodness of fit of the configural models was evaluated through descriptive fit indices and residual values: CFI, TLI, SRMR, and RMSEA following the criteria by Isac et al. (2019), Kline (2016), and Sideridis et al. (2018) so that (a) values above .90 for CFI and TLI, and (b) values below 0.06 for SRMR and RMSEA indicate good fit.

On the other hand, the change in model fit between the increasingly restricted models was evaluated through chi-square difference tests for the models with equal thresholds (intercepts) and loadings however, due to the well-known sensitivity of the chi-square statistic to sample size (Svetina et al., 2020), additional model fit indices were used including the change in the comparative fit index (CFI) and the change in the root-mean-square error of approximation (RMSEA) between the increasingly constrained models. The "difftest" option in Mplus was used to obtain correct values for the chi-square difference statistics. The cutoff values used as indicators of ***non-invariance*** were:

- For configural invariance: $MSEA \leq 0.08,\ CFI \geq 0.90.$
- For metric invariance: $\Delta RMSEA \leq 0.05,\ \Delta CFI \geq -0.004$ and significant X^2.
- For scalar invariance: $\Delta RMSEA \leq 0.01,\ \Delta CFI \geq -0.004$ and significant X^2 (Svetina et al., 2020).

All the analyses were conducted in Mplus (Version 8.1).

3.3.2.2 The Alignment Optimization. A configural model was defined for each of the measures where the item loadings and intercepts were freely estimated and the factor mean and variances were set to 0 and 1, respectively using the FIXED option available in Mplus. Moreover, to control for the nesting of students within schools, the TYPE= COMPLEX option in Mplus was applied.

Then, the alignment procedure was implemented where the group factor mean (α_g) and factor variance (ψ_g) are chosen through the simplicity function (see equation 36) to minimize the amount of measurement noninvariance. The method was implemented in two steps:

1. Estimation of the configural model where the loadings and intercepts were freely estimated across groups while keeping the factor means and factor variances fixed to 0 and 1, respectively across the groups.
2. Alignment optimization where the factor means and variances were freely estimated and their values were chosen to minimize the total amount of noninvariance so that they became aligned (invariant) across the groups, through the simplicity function F for every pair of groups and every intercept and factor loading using the component loss function f from exploratory factor analysis (EFA) rotations where

$$F = \sum_p \sum_{j_1<j_2} w_{j_1,j_2} f\left(\lambda_{pj_1} - \lambda_{pj_2}\right) + \sum_p \sum_{j_1<j_2} w_{j_1,j_2} f\left(\nu_{pj_1} - \nu_{pj_2}\right) \quad (44)$$

In this scenario, the nonidentified model where the factor means, and variances were added to the configural model is now identified by adding the simplicity requirement. The procedure provides an R^2 measure to indicate how much of the configural parameter variation across groups can be explained by the variation the factor means and variances so that high values indicate a high degree of measurement invariance.

The measurement models that were analyzed through the alignment optimization procedure for the exposure to bullying, sense of belonging at school, and reading scales are shown in figures 7, 8, and 9, respectively.

Figure 7

Bullying Measurement Model for Alignment Optimization

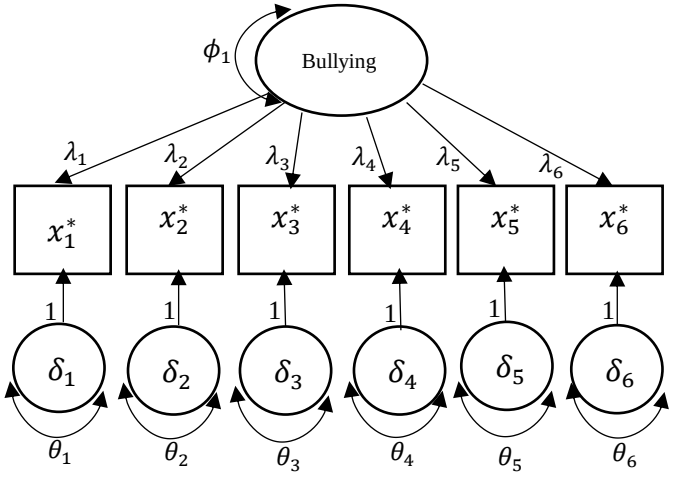


Figure 8

Sense of Belonging at School Measurement Model for Alignment Optimization

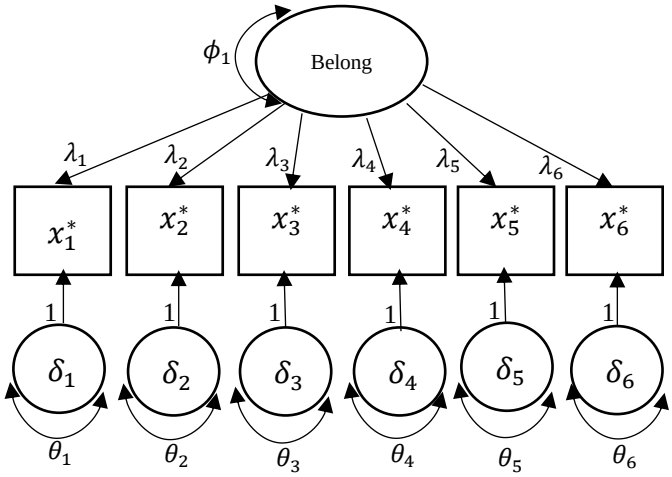
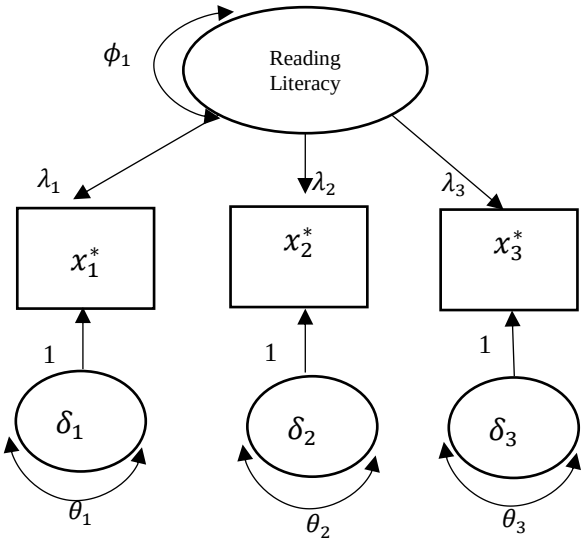


Figure 9

Reading Literacy Measurement Model for Alignment Optimization



3.3.3 Evaluation of the Relationship between the Non-cognitive Measures and the Performance on Reading literacy

A structural equation model (shown in Figure 10) was tested to evaluate the extent to which the non-cognitive measures predict test performance on reading literacy.

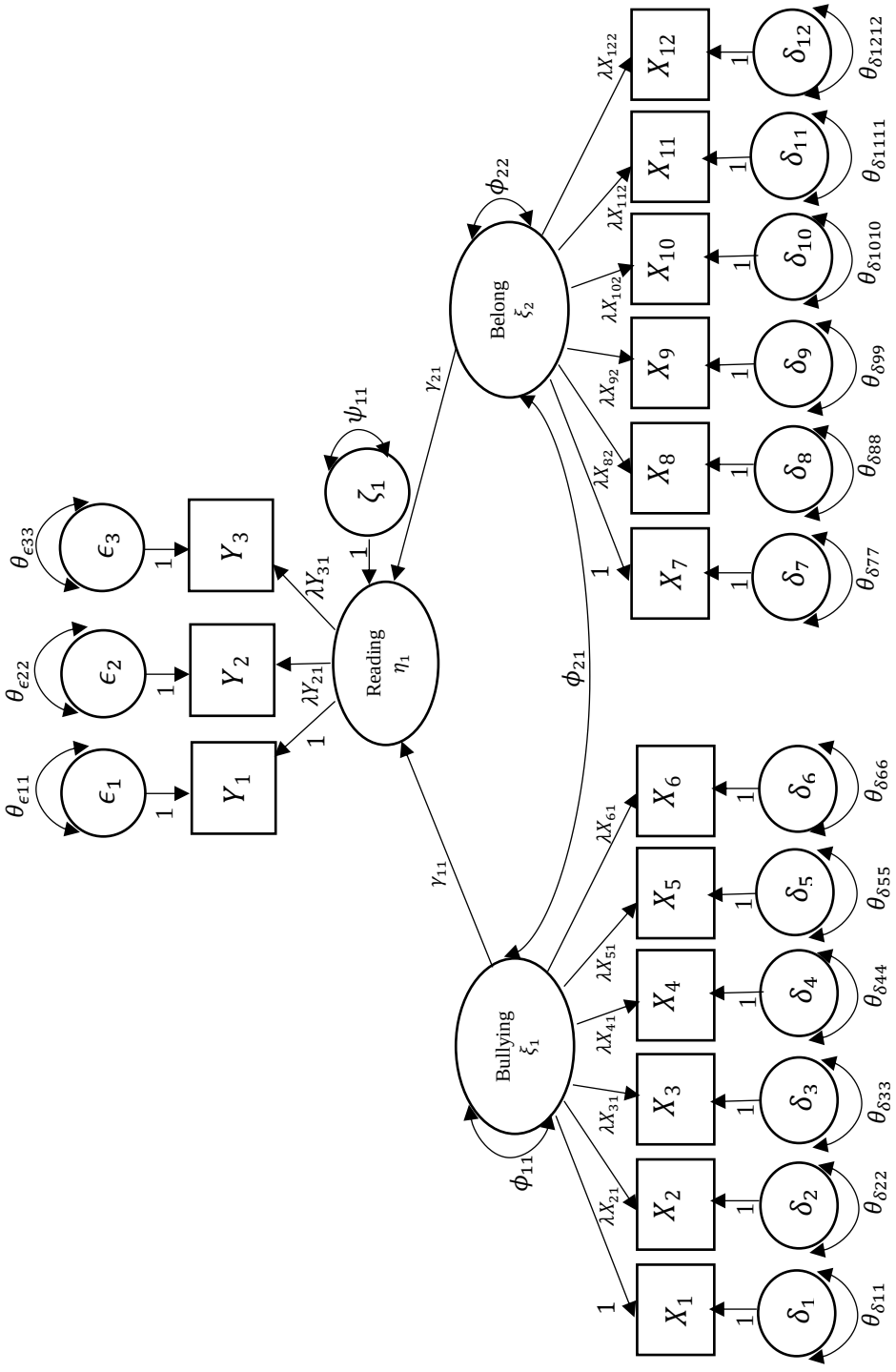
The discrepancies between the data and hypothesized variance-covariance matrices were evaluated through an omnibus chi-square test and model fit was assessed through the comparative fit index (CFI), the Tucker-Lewis index (TLI), SRMR, and RMSEA (Kline, 2016; Sideridis et al., 2018).

In terms of cutoff criteria for the indices, values of RMSEA between 0.080 and 0.010 were considered as indicators of acceptable fit and values of CFI and TLI between 0.9 and 0.95 were also considered as indicators of acceptable model fit (Isac et al., 2019).

The analyses were conducted in Mplus (Version 8.1).

Figure 10

Latent Structural Regression Model of Reading Literacy and Non-Cognitive Measures



CHAPTER

IV RESULTS

This chapter includes the results from the evaluation of measurement invariance and it is organized as follows: (a) *descriptive analyses* of the sample based on socio-demographic variables across countries, (b) descriptive analyses per scale where the distribution of the latent constructs is provided across countries and per immigration status of the students, (c) evaluation of measurement invariance through multiple group confirmatory factor analysis per scale, (d) evaluation of measurement invariance through the alignment method per scale, and (e) evaluation of the relationship between the non-cognitive measures and the performance on the reading literacy scale.

4.1 Descriptive Analyses

4.1.1 Sample

The sample was described in terms of the gender and economic, social, and cultural status (ESCS) of the students. Table 6 shows the distribution of gender by immigration status per country.

Table 6

Frequency of Gender per Immigration Status across Countries

	Native		Immigrant		
Country	**Female**	**Male**	**Female**	**Male**	**Total**
Slovenia	2076	2238	183	179	4676
Netherlands	1542	1550	195	194	3481
Germany	826	910	204	218	2158
Estonia	2139	1980	206	256	4581
Brunei Darussalam	2041	1983	208	211	4443
Croatia	2375	2221	245	223	5064
Serbia	2002	1886	249	180	4317
France	1800	1821	254	270	4145
Greece	2373	2190	262	264	5089
Costa Rica	2685	2535	288	288	5796
Norway	2263	2154	311	282	5010
Italy	3590	3740	347	373	8050
Ireland	1735	1769	380	326	4210
Sweden	1936	1775	384	377	4472
United States	1703	1685	441	449	4278
Kazakhstan	6708	6549	480	533	14270
Austria	2076	1944	510	466	4996
Switzerland	1047	1142	517	587	3293
Belgium	2932	2763	547	542	6784
New Zealand	1906	1598	577	617	4698
Denmark	2325	2358	597	473	5753
Singapore	2323	2434	710	708	6175
United Kingdom	5051	4724	766	662	11203
Hong Kong	1707	1653	983	1058	5401
Luxembourg	999	968	1074	1143	4184
Macao	648	710	1152	1135	3645
Australia	3620	3677	1327	1373	9997
Spain	11097	11028	1380	1392	24897
Canada	6853	6278	2292	2281	17704
Qatar	2430	1641	3243	2888	10202
United Arab Emirates	3820	3119	4311	4093	15343

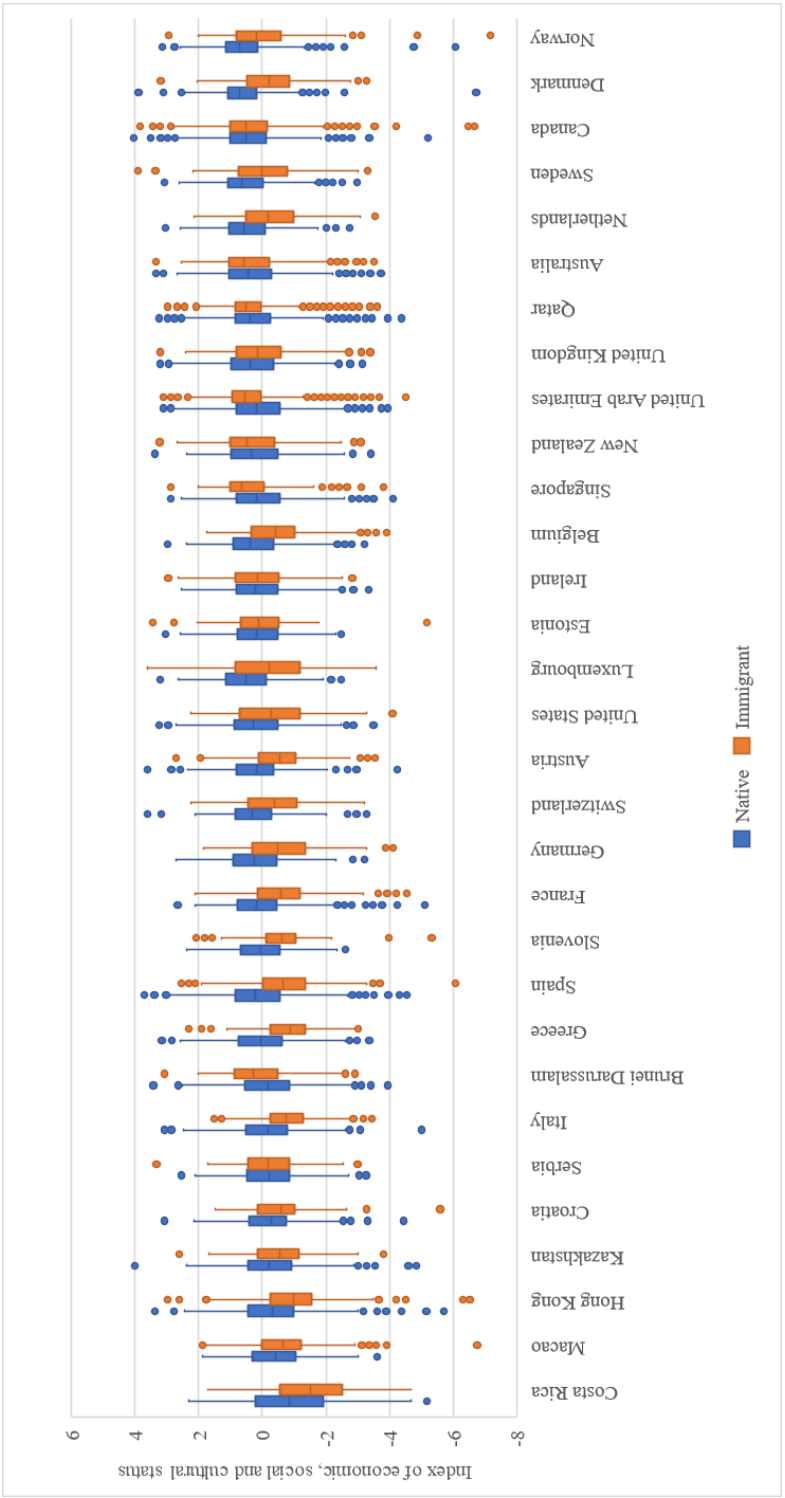
According to Table 6, the distribution of gender seems similar across countries so that approximately half of the students were identified as male, and this distribution holds across immigrant and native students. The frequency of female students with respect to male students seems to be higher among native students in Macao, Serbia, Brunei

Darussalam, Sweden, Estonia, New Zealand, Austria, Norway, Croatia, Greece, Hong Kong, Belgium, Qatar, United Kingdom, United Arab Emirates, and Canada. Regarding the subpopulation of immigrant students, the frequency of female students is higher with respect to male students in Macao, Ireland, Serbia, Sweden, Austria, Norway, Croatia, Qatar, United Kingdom, and United Arab Emirates.

Figure 8 shows the distribution of ESCS between native and immigrant students across countries. The ESCS PISA index measures the access students have to family resources in terms of financial, social and cultural capital, which determines the social position of the students' household, and it is usually used as a good approximation of inequality of opportunity among students. The index is a weighted average of three indices: parental education, parental occupation, and household possessions, and it is normalized to have a mean of zero and standard deviation of one (Avvisati, 2020). As shown in Figure 8, the median value of the ESCS index among *native* students tends to be around zero or above for most countries except Costa Rica, Croatia, Hong Kong, Italy, Kazakhstan, Macao, and Serbia where the median value is around -1. However, the differences do not appear to be large, and most values seem to fall between 3 and -4.

Figure 11

Distribution of ESCS Index per Status as Immigrant across Countries



The variability of the ESCS index seems more noticeable among immigrant students where the means for most countries tend to fall below zero with a few exceptions including Canada, Qatar, Singapore, and United Arab Emirates.

On the other hand, the variability of the median ESCS index *within countries* is more noticeable between immigrant and native students in Austria (-0.5 and 0.17, respectively), Belgium (-0.4 and 0.4), Denmark (-0.2 and 0.7), France (-0.6 and 0.2), Germany (-0.5 and 0.2), Greece (-0.9 and 0.1), Luxembourg (-0.2 and 0.5), Netherlands (-0.2 and 0.6), Slovenia (-0.6 and 0), Spain (-0.7 and 0.2), Switzerland (-0.4 and 0.3), and United States (-0.3 and 0.3). The median values for the remaining countries are similar for both immigrant and native students.

4.1.2 Non-cognitive Measures

The psychometric properties of the non-cognitive measures were described at the item level where each measure included six categorical items with four answer choices. Results for the bullying scale are shown next followed by the sense of belonging at school scale.

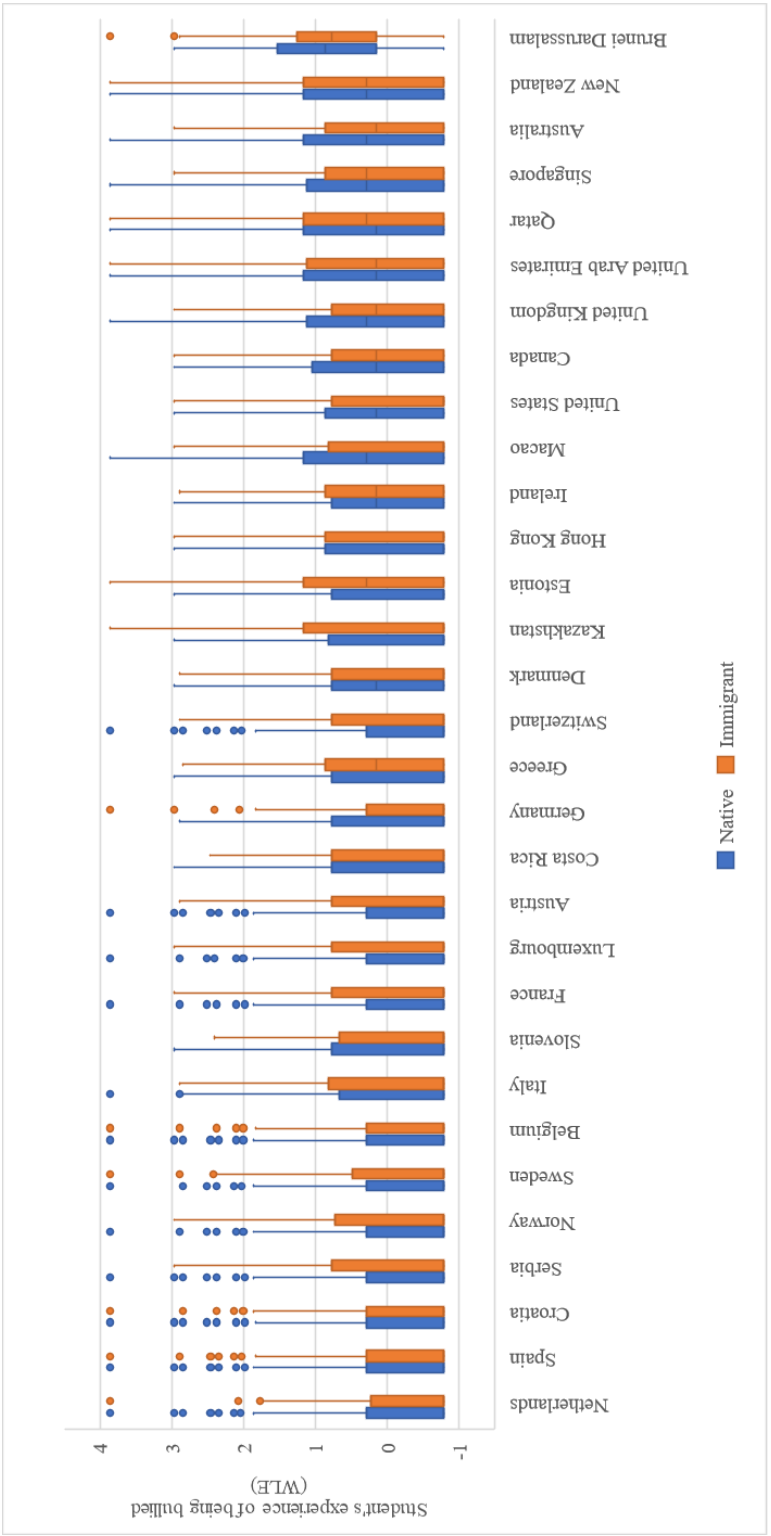
4.1.2.1 Bullying. As previously mentioned, when presented with this scale, students are asked to indicate how often (never or almost never, a few times a year, a few times a month, once a week or more) they have had the experiences expressed in the items during the 12 months prior to the test.

PISA provides an index of bullying with an average of zero and standard deviation of 1 so that positive values indicate more exposure to bullying at school than the average student in OECD countries while negative values indicate less exposure to bullying. Specifically, a value greater than 1.51 suggests frequent exposure to bullying.

The distribution of the bullying index across countries is shown in Figure 12. According to Figure 12, the median value of the index appears to be similar between immigrant and native students across countries except for Denmark, Macao, and the United States where the median values were higher (above zero) suggesting higher levels of bullying for native students than immigrant students (values closer to -1), and Estonia and Greece where the median index was higher (above zero) for immigrant students. The country with the highest median index of bullying across immigrant and native students was Brunei Darussalam followed by New Zealand, and Singapore.

Figure 12

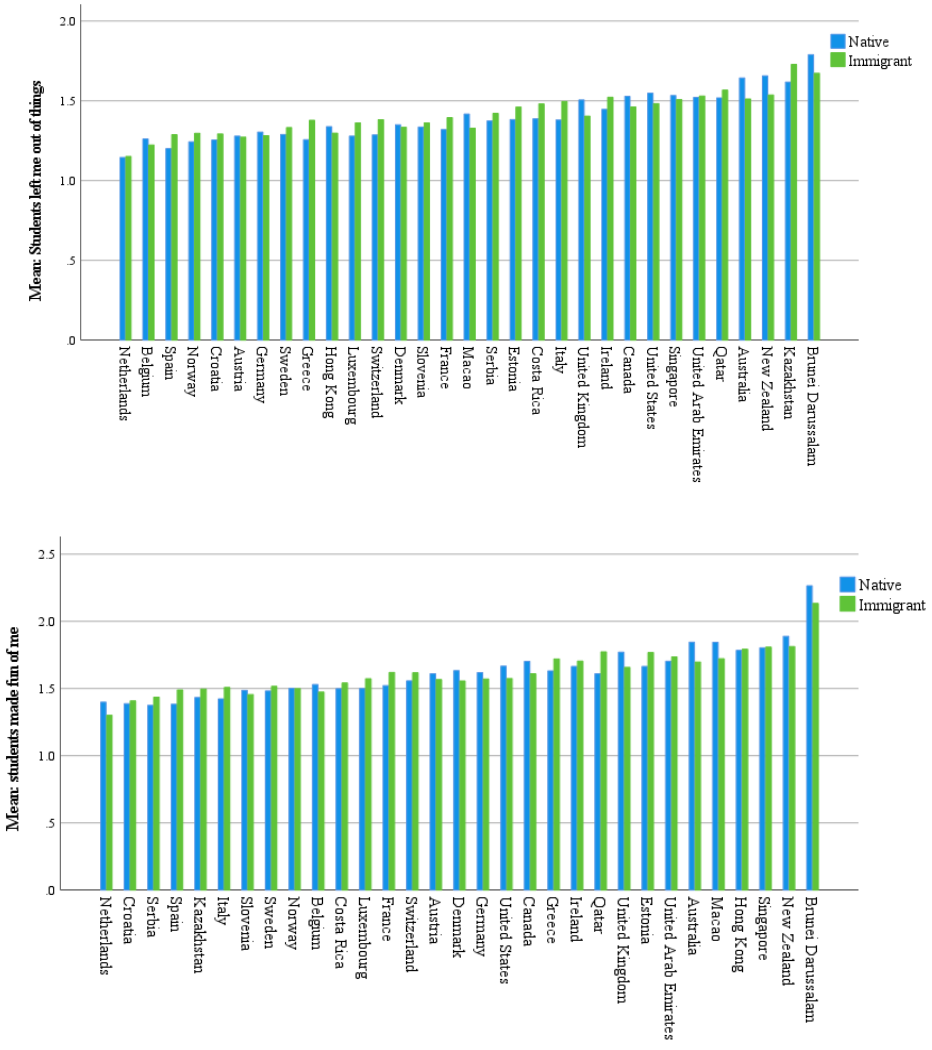
Distribution of Bullying Index per Immigration Status across Countries

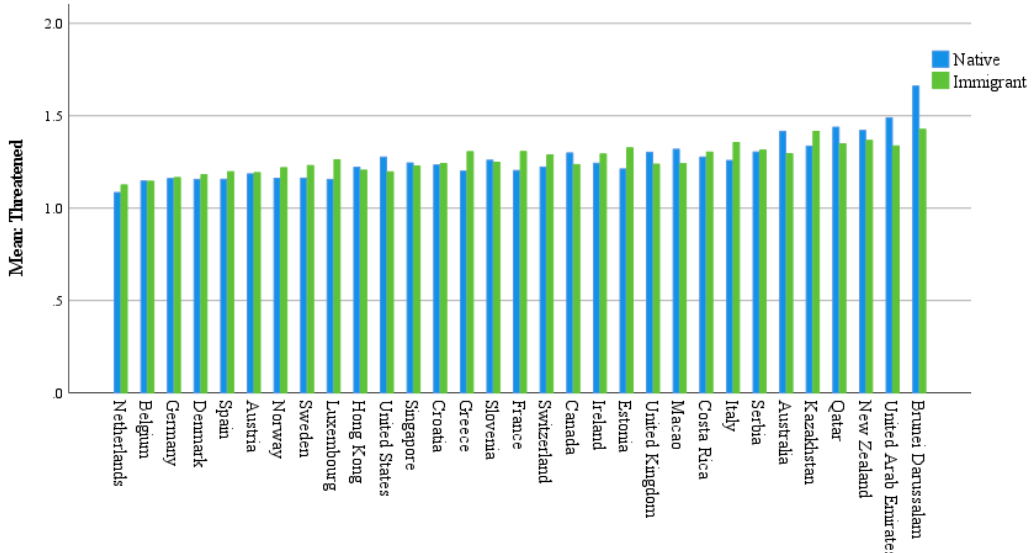


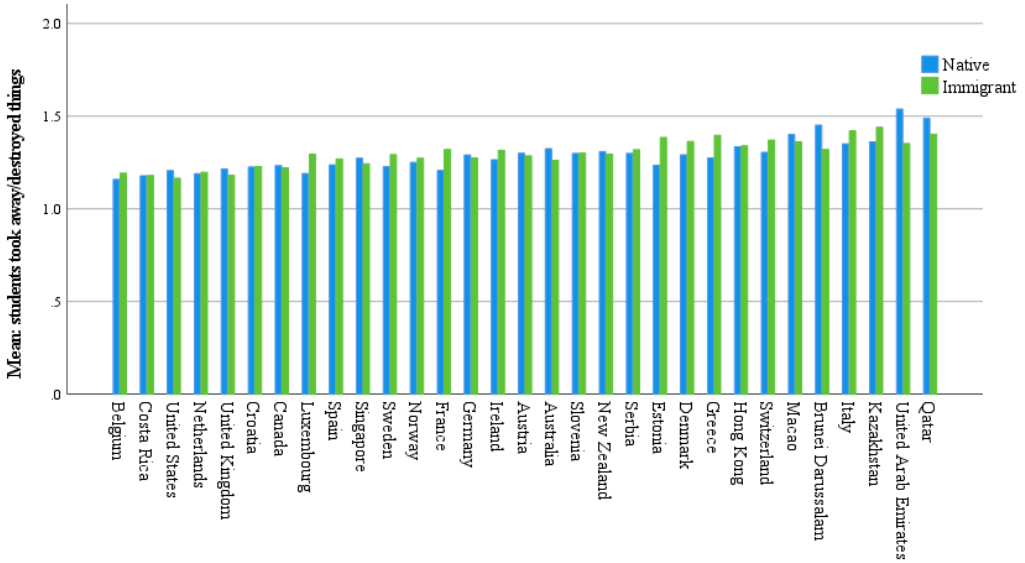
Mean item scores were described next per country and across students. Results are shown in Figure 13. The mean scores for item 1 fall between 1.2 (Belgium and Netherlands) and 1.7 (Brunei Darussalam and Kazakhstan) suggesting that most students across countries tend to select the lower response categories that indicate low levels of exposure to the bullying indicator. In terms of the subpopulations of immigrant and native students, results show that the mean values for item 1 (left out of things) tend to be similar between the groups of students across countries except for Australia, Brunei Darussalam, and New Zealand where the differences are more noticeable so that native students report higher values.

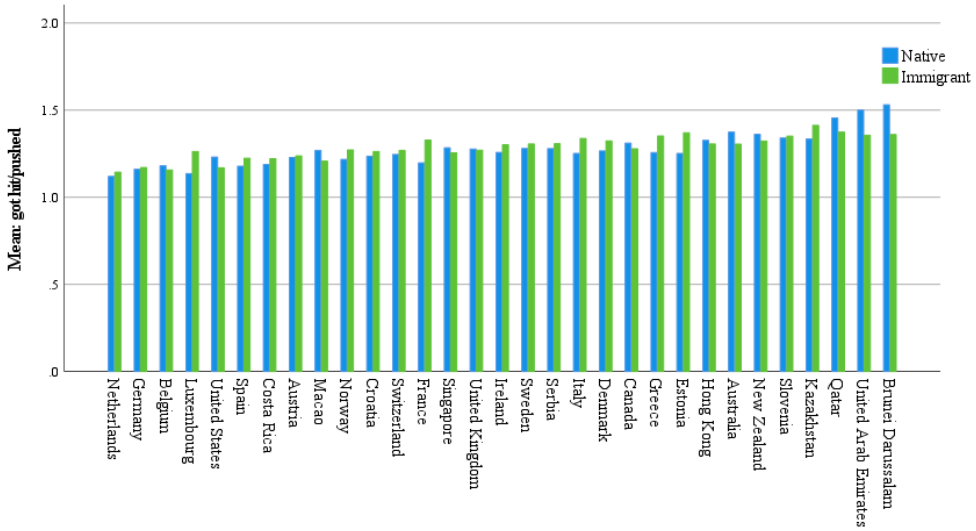
Figure 13

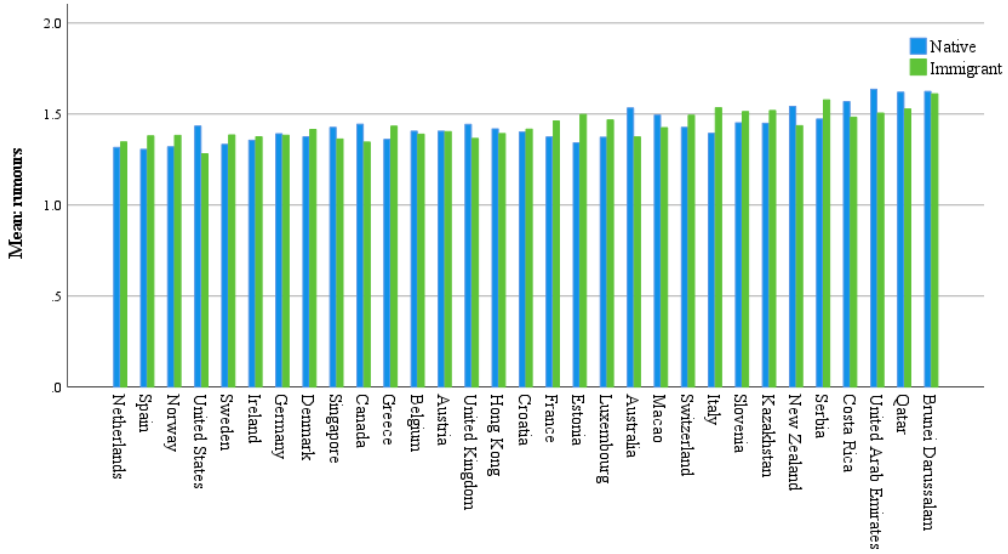
Mean Item Scores Bullying Scale per Immigration Status across Countries




Mean: Threatened
2.0
1.5
1.0
.5
.0
Native
Immigrant
Netherlands
Belgium
Germany
Denmark
Spain
Austria
Norway
Sweden
Luxembourg
Hong Kong
United States
Singapore
Croatia
Greece
Slovenia
France
Switzerland
Canada
Ireland
Estonia
United Kingdom
Macao
Costa Rica
Italy
Serbia
Australia
Kazakhstan
Qatar
New Zealand
United Arab Emirates
Brunei Darussalam


Mean: students took away/destroyed things
2.0
1.5
1.0
.5
.0
Native
Immigrant
Belgium
Costa Rica
United States
Netherlands
United Kingdom
Croatia
Canada
Luxembourg
Spain
Singapore
Sweden
Norway
France
Germany
Ireland
Austria
Australia
Slovenia
New Zealand
Serbia
Estonia
Denmark
Greece
Hong Kong
Switzerland
Macao
Brunei Darussalam
Italy
Kazakhstan
United Arab Emirates
Qatar


Mean: got hit/pushed
2.0
1.5
1.0
.5
.0
Native
Immigrant
Netherlands
Germany
Belgium
Luxembourg
United States
Spain
Costa Rica
Austria
Macao
Norway
Croatia
Switzerland
France
Singapore
United Kingdom
Ireland
Sweden
Serbia
Italy
Denmark
Canada
Greece
Estonia
Hong Kong
Australia
New Zealand
Slovenia
Kazakhstan
Qatar
United Arab Emirates
Brunei Darussalam


Mean: rumours
2.0
1.5
1.0
.5
.0
Native
Immigrant
Netherlands
Spain
Norway
United States
Sweden
Ireland
Germany
Denmark
Singapore
Canada
Greece
Belgium
Austria
United Kingdom
Hong Kong
Croatia
France
Estonia
Luxembourg
Australia
Macao
Switzerland
Italy
Slovenia
Kazakhstan
New Zealand
Serbia
Costa Rica
United Arab Emirates
Qatar
Brunei Darussalam

Regarding item 2 (students made fun of me), results show that the mean values range from 1.3 (Netherlands) to 2.1 (Brunei Darussalam) indicating that as with item 1, students tend to select the low response categories for the statement described in this item which in turn corresponds to low exposure to bullying. The mean values between native and immigrant students are also similar except for Australia, Brunei Darussalam, and Macao where the values seem higher for native students than immigrant students.

Results for item 3 (threatened) show that the mean values range from 1.1 (Belgium and Netherlands) to 1.7 (Brunei Darussalam). Like the other items, the values correspond to the lower response categories suggesting that most students tend to have low levels on this expression of bullying. In terms of immigrant and native students, results show that there seem to be differences between the two groups within the countries. For instance, in Australia, Brunei Darussalam, Canada, Macao, New Zealand, Qatar, United Arab Emirates, United Kingdom, and the United States the mean values are higher (reflecting higher perceived bullying) for native than for immigrant students whereas in Estonia, France, Ireland, Italy, Kazakhstan, Sweden, and Switzerland the values are higher for immigrant students. In general, results show variability within countries for this item.

Mean values for item 4 (took away/destroyed things) range from 1.2 (most countries) to 1.5 (United Arab Emirates). Just as with the other items, the values tend to be around the low response categories that indicate low frequency of the bullying expression stated in this item. Mean values between native and immigrant students show some variation so that immigrant students tend to have higher average exposure to bullying than their native peers (e.g., Denmark, Estonia, Greece, Italy, Kazakhstan)

except in Australia, Brunei Darussalam, Qatar, and United Arab Emirates where the most noticeable differences indicate higher mean values for native students.

Mean values for item 5 (got hit/pushed) range from 1.1 (Netherlands) to 1.4 thus, variability seems to be low for this item across countries. Differences between immigrant and native students within each country do not seem large and in most cases, the values are higher for immigrant than native students in some countries including Costa Rica, Denmark, Estonia, France, Greece, Ireland, Italy, Kazakhstan, and Luxembourg. Whereas the values are higher for native students in Australia, Brunei Darussalam, Macao, New Zealand, Qatar, United Arab Emirates and the United States.

Finally, the distribution of the mean values for item 6 (rumors) shows high variability within countries. The values range from 1.2 to 1.6 and the larger differences at the within level are in Estonia, France, Greece, Italy, Kazakhstan, Luxembourg, Macao, New Zealand, Norway, Qatar, Serbia, United Arab Emirates, and the United States. Despite these differences, the mean values in general are around the lowest response category suggesting that students in general selected the lower response choices that refer to low frequency.

Item statistics were estimated across native and immigrant students, specifically item means, and discrimination were estimated. Results are shown in Table 7.

Table 7

Item Statistics Bullying Scale per Immigration Status

Item	Native		Immigrant	
	Mean	**Discrimination**	**Mean**	**Discrimination**
3. Threatened.	1.27	0.68	1.28	0.67
5. Got hit/pushed.	1.28	0.65	1.30	0.64
4. Took/destroyed things.	1.29	0.61	1.31	0.63
1. Left out	1.41	0.56	1.45	0.55
6. Rumors.	1.43	0.62	1.44	0.61
2. Students made fun.	1.59	0.62	1.66	0.60

According to the results shown in Table 7, the item means ranged between 1.27 and 1.59, and between 1.28 and 1.66 for native and immigrant students, respectively. Items 5 (got hit/pushed) and 3 (threatened) showed the lowest mean value suggesting that students experiencing low levels of bullying are likely to select these items. Items 6 (rumors) and 2 (students made fun of me) on the other hand, showed the highest mean value among native students suggesting that students experiencing higher frequency of bullying are likely to select these items. Items 2 (students made fun of me) and 1 (left out) showed the highest mean values among immigrant students suggesting that students experiencing higher frequency of bullying tend to select these items.

Regarding item discrimination, results showed that item 3 (threatened) had the highest discrimination across native and immigrant students thus, this item best discriminates between students with high and low levels of bullying whereas item 1 (left out) showed the lowest value of discrimination. Apart from these items, most items in the scale showed a discrimination value above 0.60 across the subpopulations of students.

4.1.2.2 Sense of Belonging at School. The six items from the sense of belonging at school scale were also described in terms of their psychometric properties. In this

scale, students were asked to indicate their level of agreement with each of the statement described in each item using four response categories: strongly disagree, disagree, agree, and strongly agree.

Given that (a) the answer choices were presented to students in the following order

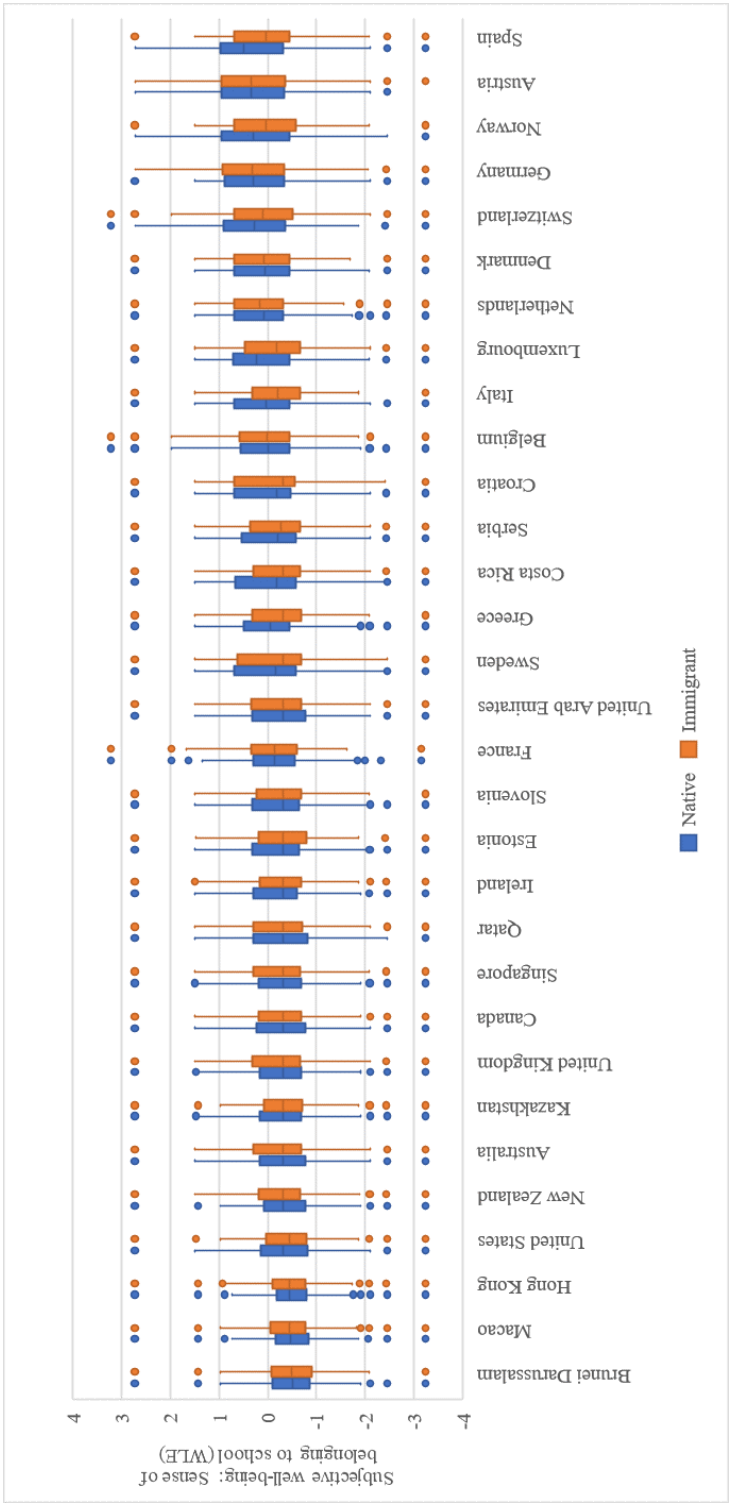
1. Strongly agree
2. Agree
3. Disagree
4. Strongly disagree

and (b) the wording of the items was not consistent so that items 2 (“I make friends easily at school”), 3 (“I feel like I belong at school”), and 5 (“Other students seem to like me”) were positively worded while the remaining were negatively worded, items 2, 3, and 5 were reverse coded to maintain the original order of the answer choices and guarantee that the highest choice would indicate higher sense of belonging at school.

Like the bullying scale, PISA provides an index of sense of belonging that has a mean of 0 and standard deviation of 1. Positive values indicate that students have a strong sense of belonging at school than the average students in OECD countries so that they are likely to feel accepted, respected, and supported in their social context at school. The distribution of this index across countries is shown in Figure 14.

Figure 14

Distribution of Sense of Belonging Index per Immigration Status across Countries

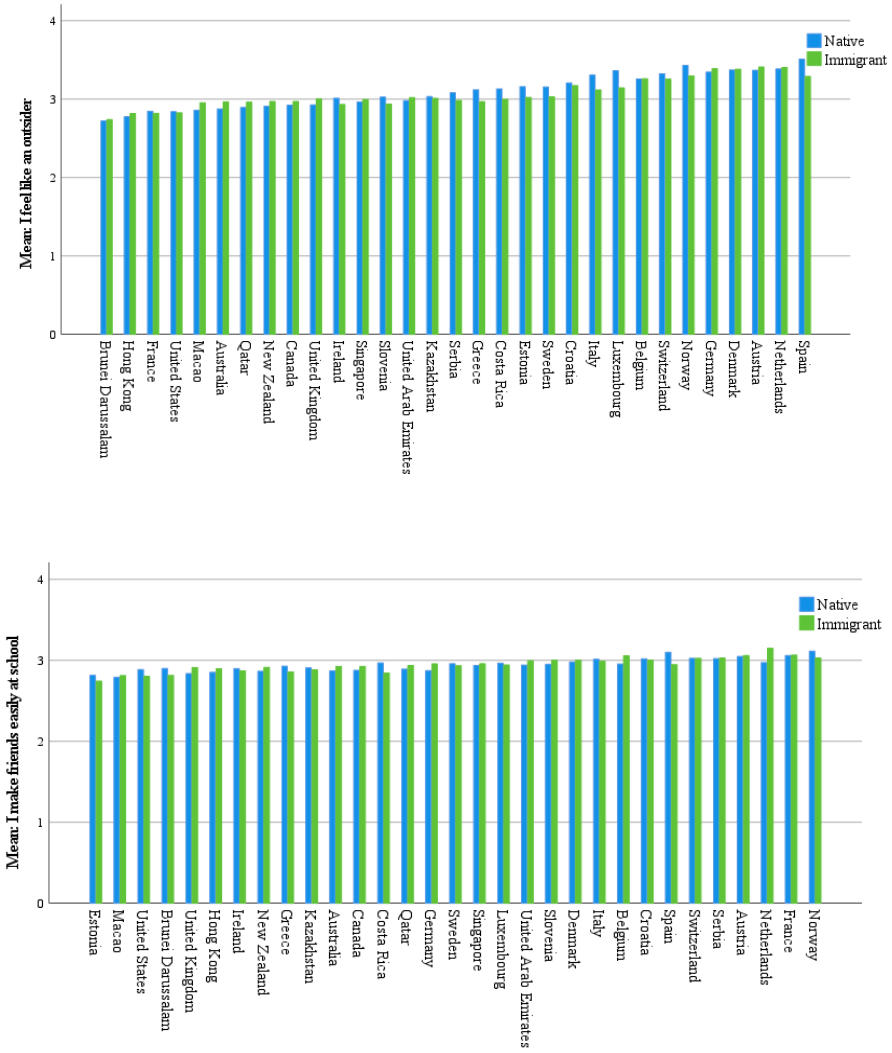


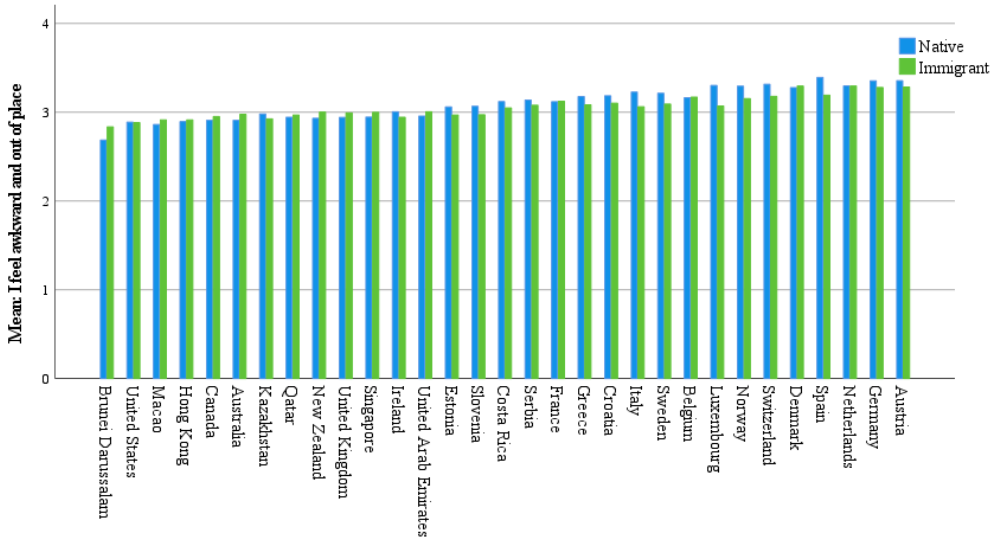
According to Figure 14, the median values of the index are between -1 and 1 across countries and the values seem to be similar across immigrant and native students, except for Spain where the median differences are more noticeable. The countries with the highest median values include Austria, Germany, and Spain suggesting that students within these countries report higher levels of the latent construct than the remaining countries. Brunei Darussalam, Hong Kong, and Macao on the other hand, showed the lowest median values.

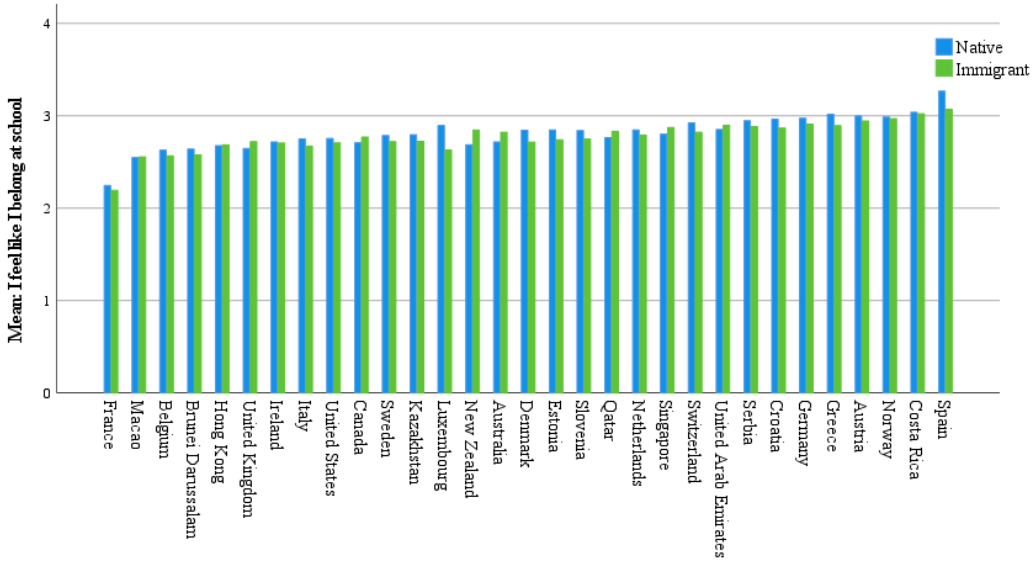
Mean item scores were described next per country and across students in Figure 15. The mean score for item 1 ranges from 2.7 (Brunei Darussalam) to 3.5 (Spain). Countries with item mean above 3 for both immigrant and native students, include Austria, Belgium, Croatia, Denmark, Germany, Italy, Luxembourg, Netherlands, Norway, Spain, and Switzerland. This finding suggests that students within these countries are likely to select the higher response categories which indicate higher levels of sense of belonging at school. The countries with the lowest mean included Brunei Darussalam and Hong Kong and the largest mean differences between immigrant and native students seem to be more noticeable in Spain, Italy, and Luxembourg so that native students obtained a higher mean than their immigrant peers. In general, the means between immigrant and native students within each country seemed similar.

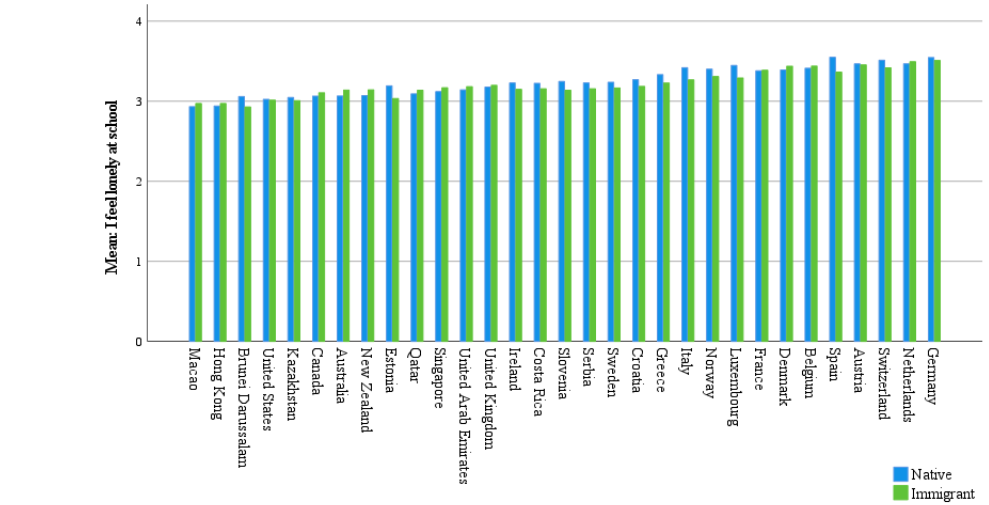
Figure 15

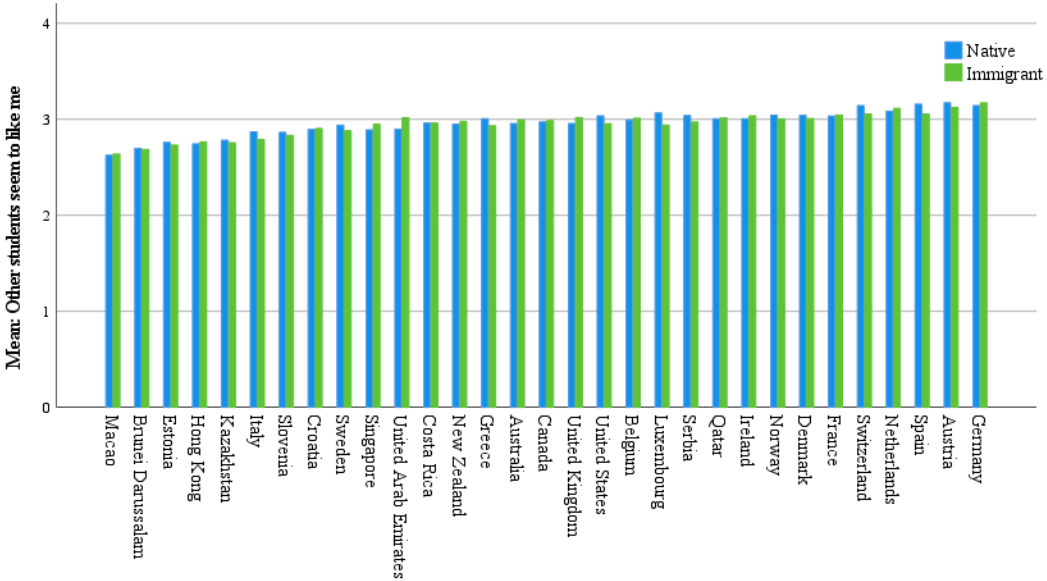
Mean Item Scores Sense of Belonging at School Scale per Immigration Status across Countries




Mean: I feel awkward and out of place
0
1
2
3
4
Native
Immigrant
Brunei Darussalam
United States
Macao
Hong Kong
Canada
Australia
Kazakhstan
Qatar
New Zealand
United Kingdom
Singapore
Ireland
United Arab Emirates
Estonia
Slovenia
Costa Rica
Serbia
France
Greece
Croatia
Italy
Sweden
Belgium
Luxembourg
Norway
Switzerland
Denmark
Spain
Netherlands
Germany
Austria


Mean: I feel like I belong at school
0
1
2
3
4
Native
Immigrant
France
Macao
Belgium
Brunei Darussalam
Hong Kong
United Kingdom
Ireland
Italy
United States
Canada
Sweden
Kazakhstan
Luxembourg
New Zealand
Australia
Denmark
Estonia
Slovenia
Qatar
Netherlands
Singapore
Switzerland
United Arab Emirates
Serbia
Croatia
Germany
Greece
Austria
Norway
Costa Rica
Spain


Mean: I feel lonely at school
0
1
2
3
4
Macao
Hong Kong
Brunei Darussalam
United States
Kazakhstan
Canada
Australia
New Zealand
Estonia
Qatar
Singapore
United Arab Emirates
United Kingdom
Ireland
Costa Rica
Slovenia
Serbia
Sweden
Croatia
Greece
Italy
Norway
Luxembourg
France
Denmark
Belgium
Spain
Austria
Switzerland
Netherlands
Germany
Native
Immigrant


Mean: Other students seem to like me
0
1
2
3
4
Native
Immigrant
Macao
Brunei Darussalam
Estonia
Hong Kong
Kazakhstan
Italy
Slovenia
Croatia
Sweden
Singapore
United Arab Emirates
Costa Rica
New Zealand
Greece
Australia
Canada
United Kingdom
United States
Belgium
Luxembourg
Serbia
Qatar
Ireland
Norway
Denmark
France
Switzerland
Netherlands
Spain
Austria
Germany

The mean for item 2 (“I make friends easily at school”) seems to be consistent across countries and within countries around 2.9 suggesting that most students tend to choose the higher response choices that indicate higher level of sense of belonging at school. The average of item 3 (“I feel like I belong at school”) ranges from 2.2 (France) and 3.1 (Spain). In general, the item mean is around 2.7 suggesting that students tend to select response categories that reflect middle level of sense of belonging at school. Moreover, item means seem to be similar between immigrant and native students; the most noticeable difference can be seen in Luxembourg and Spain.

The means for item 4 (I feel awkward and out of place in my school”) ranges from 2.7 (Brunei Darussalam) to 3.4 (Austria and Spain) and in general, the mean values tend to be above 3 for most countries. This finding indicates that most students tend to select the high response choices that represent a high level of the construct when presented with this item. The means within countries also tend to be similar across native and immigrant students except for Luxembourg and Spain that show a larger difference at the within level than the rest of the countries.

The means for item 5 (“other students seem to like me”) do not show a large variation across countries and most values are around 3 suggesting that students across most countries tend to select the “high” response categories which in turn indicate high values of the latent construct. Brunei Darussalam, Estonia, Hong Kong, and Macao showed the lowest means. The mean values across immigrant and native students within each country seem to be similar implying that this item might be interpreted in a similar way across countries and students.

Regarding item 6 (“I feel lonely at school”), results show that most of the mean values are above 3 except for Hong Kong and Macao indicating that most students across all the countries tend to select the highest answer choices that correspond to high levels of sense of belonging at school. The mean values are similar across native and immigrant students within each country and the most noticeable differences (although not large) are among students living in Estonia, Italy, Luxembourg, and Spain.

Item statistics were estimated next across immigrant and native students. Results are shown in Table 8. Results show that item means ranged from 2.85 to 3.25 among native students and from 2.80 to 3.18 among immigrant students. In general, item means suggest that students agree with the statements in this scale which in turn implies that most students have a high level of sense of belonging at school. On the other hand, the values for item discrimination range from 0.44 to 0.56 among native students and from 0.43 to 0.54 among immigrant students. Item 6 (“I feel lonely at school”) showed the highest discrimination and item 3 (“I feel like I belong at school”), the lowest. Therefore, item 6 could be used to discriminate between students with high and low levels of sense of belonging at school. In general, the discrimination values for this scale were low which was expected given the low variation in the values of item difficulty.

Table 8

Item Statistics Sense of Belonging at School Scale per Immigration Status

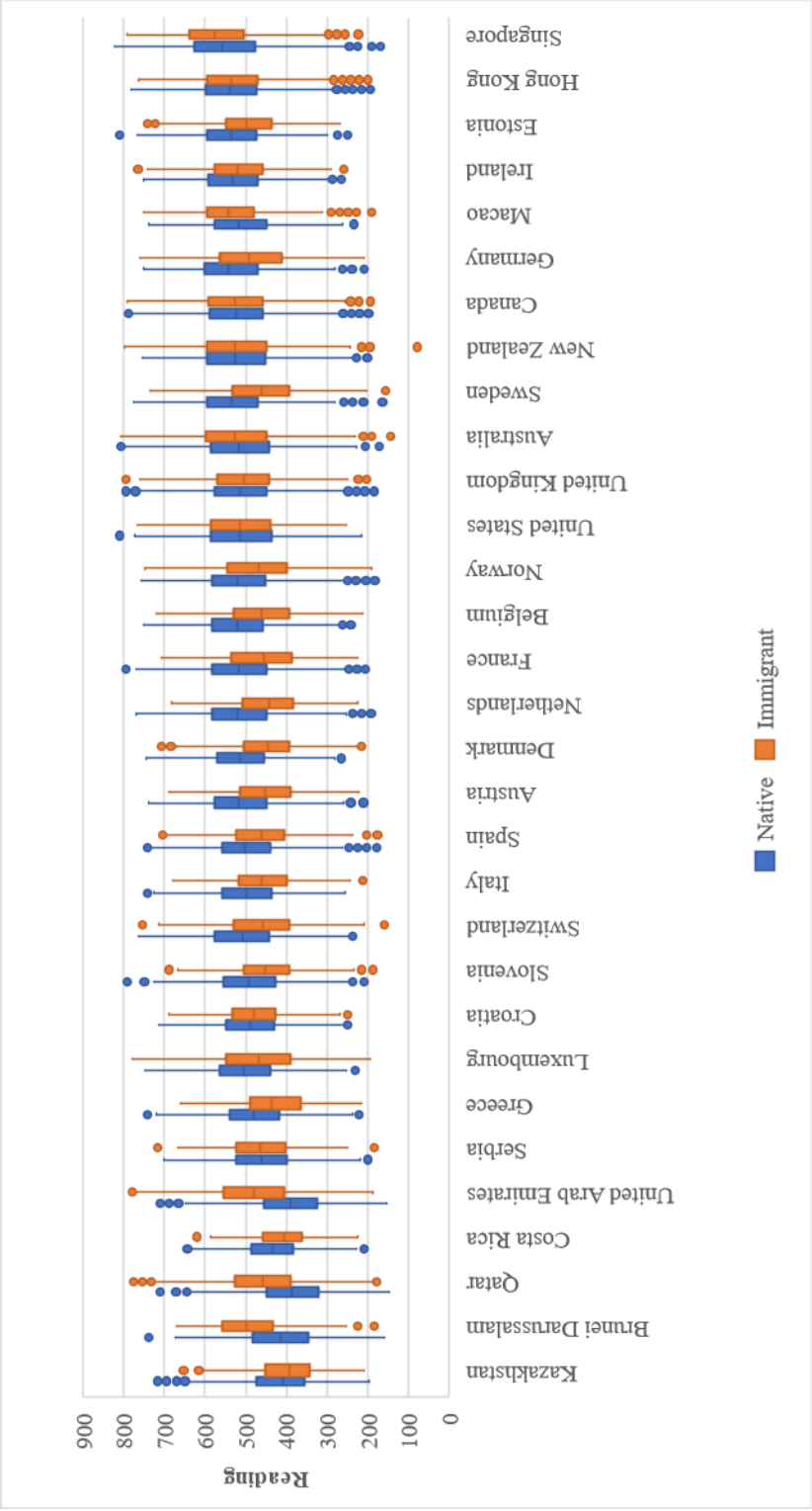
Item	Native		Immigrant	
	Mean	**Discrimination**	**Mean**	**Discrimination**
3. Feel belong.	2.85	0.44	2.80	0.43
2. Make friends easily.	2.95	0.48	2.94	0.47
5. Other students like me.	2.97	0.45	2.96	0.43
4. Feel awkward/out of place.	3.10	0.52	3.02	0.48
1. Feel like outsider.	3.12	0.50	3.04	0.48
6. Feel lonely.	3.25	0.56	3.18	0.54

4.1.3 Cognitive Measure

4.1.3.1 Reading Literacy. Descriptive analyses were conducted on the reading literacy measure at the test level. To do so, the ten plausible values reported for reading literacy were used along with the plausible values for the three subscales: locate information, understand, and evaluate and reflect.

Figure 16

Summary Reading Literacy per Country and Immigrant Status



According to Figure 16, the median scores for reading literacy fall between 300 and 600 across all countries. There distribution of the reading literacy index varies across countries for instance, Singapore showed the highest median value and Qatar along with the United Arab Emirates showed the lowest.

Moreover, there are noticeable differences in the median values within countries across native and immigrant students where immigrant students tend to score lower than their native peers in most countries except in Brunei Darussalam, Macao, Qatar, Singapore, and the United Arab Emirates where immigrant students scored higher. Additional details are shown in Table 9.

Table 9

Summary Reading Scores per Country and Immigration Status

	Native					Immigrant					Mean
Country	Mean	SD	Median	Max.	Min.	Mean	SD	Median	Max.	Min.	**diff.**
Qatar	387.6	90.8	386.9	722.2	148.3	458.9	98.5	459.3	774.9	178.7	**-71.3**
Un. Arab Em.	392.7	92.3	389.2	714.9	153.1	478.9	102.2	481.0	792.3	187.3	**-86.2**
Kazakhstan	416.7	84.6	407.6	717.2	198.2	399.8	77.4	391.8	662.8	210.3	16.9
Brunei Daruss.	416.8	95.1	414.7	737.7	158.6	490.4	92.3	499.4	669.4	183.6	**-73.6**
Costa Rica	435.7	73.3	433.5	649.0	209.9	410.7	70.1	405.9	621.2	223.8	25.0
Serbia	460.1	88.3	460.9	701.7	200.8	463.4	84.9	463.1	716.0	185.3	-3.3
Greece	476.9	87.6	479.5	742.4	190.1	430.5	89.8	435.2	659.6	215.2	46.3
Croatia	488.0	82.8	488.2	712.5	232.7	479.3	81.1	479.7	687.3	249.2	8.7
Slovenia	491.2	88.4	494.0	790.3	208.9	447.2	84.5	450.7	687.6	186.7	44.0
Italy	495.4	87.0	499.3	754.4	219.0	456.7	90.6	460.3	679.1	211.7	38.7
Spain	496.9	84.3	500.9	752.5	178.9	463.2	83.4	463.0	713.1	176.4	33.7
Luxembourg	500.2	92.1	504.1	748.5	230.5	470.1	107.8	469.1	777.6	193.5	30.1
Switzerland	506.7	93.1	509.4	764.8	237.4	460.5	100.3	458.6	753.7	159.9	46.2
United States	509.1	100.9	513.6	810.5	217.1	510.9	102.2	513.9	767.7	254.4	-1.8
Austria	510.4	88.7	516.2	737.6	210.8	454.0	85.5	453.6	688.9	221.8	56.4
Australia	511.6	102.4	516.7	809.2	172.5	519.7	106.1	528.0	808.2	143.2	-8.1
Netherlands	511.9	95.1	519.6	768.5	191.9	445.2	89.9	443.6	681.9	224.4	**66.7**
United Kingdom	512.1	90.6	514.9	793.4	184.6	501.3	93.0	505.3	794.6	202.8	10.8
Denmark	512.2	83.7	515.5	745.9	230.4	448.7	81.4	445.7	707.2	214.5	**63.5**
Macao	512.7	90.2	518.2	738.8	233.9	533.9	86.0	541.0	749.4	191.7	-21.2
France	512.7	92.3	518.5	793.4	205.6	460.8	97.9	455.1	705.9	225.0	51.9
Norway	514.5	95.6	520.7	756.8	182.9	468.4	98.2	468.7	748.9	190.2	46.1
Belgium	517.6	88.9	521.1	751.0	242.1	462.2	94.1	461.1	720.1	212.4	55.4
New Zealand	520.1	97.9	525.9	753.6	201.6	519.1	105.2	526.9	798.5	79.3	1.0
Canada	521.3	93.3	525.2	789.4	198.5	524.2	95.1	527.9	792.3	194.8	-2.9

Ireland	528.9	85.3	532.7	751.1	241.5	517.2	86.4	519.8	764.9	259.5	11.7
Sweden	530.0	92.2	536.4	774.6	164.5	462.5	101.7	462.0	734.3	157.1	**67.4**
Hong Kong	531.0	91.9	540.7	783.1	194.8	528.4	92.9	537.2	762.2	199.5	2.6
Estonia	533.3	87.0	535.6	810.0	250.9	494.3	86.5	497.9	742.6	267.6	39.0
Germany	533.9	91.7	541.4	749.8	208.7	488.3	105.8	492.4	761.2	209.3	45.7
Singapore	547.3	106.5	557.3	823.4	169.9	565.7	99.9	577.2	791.8	223.6	-18.4

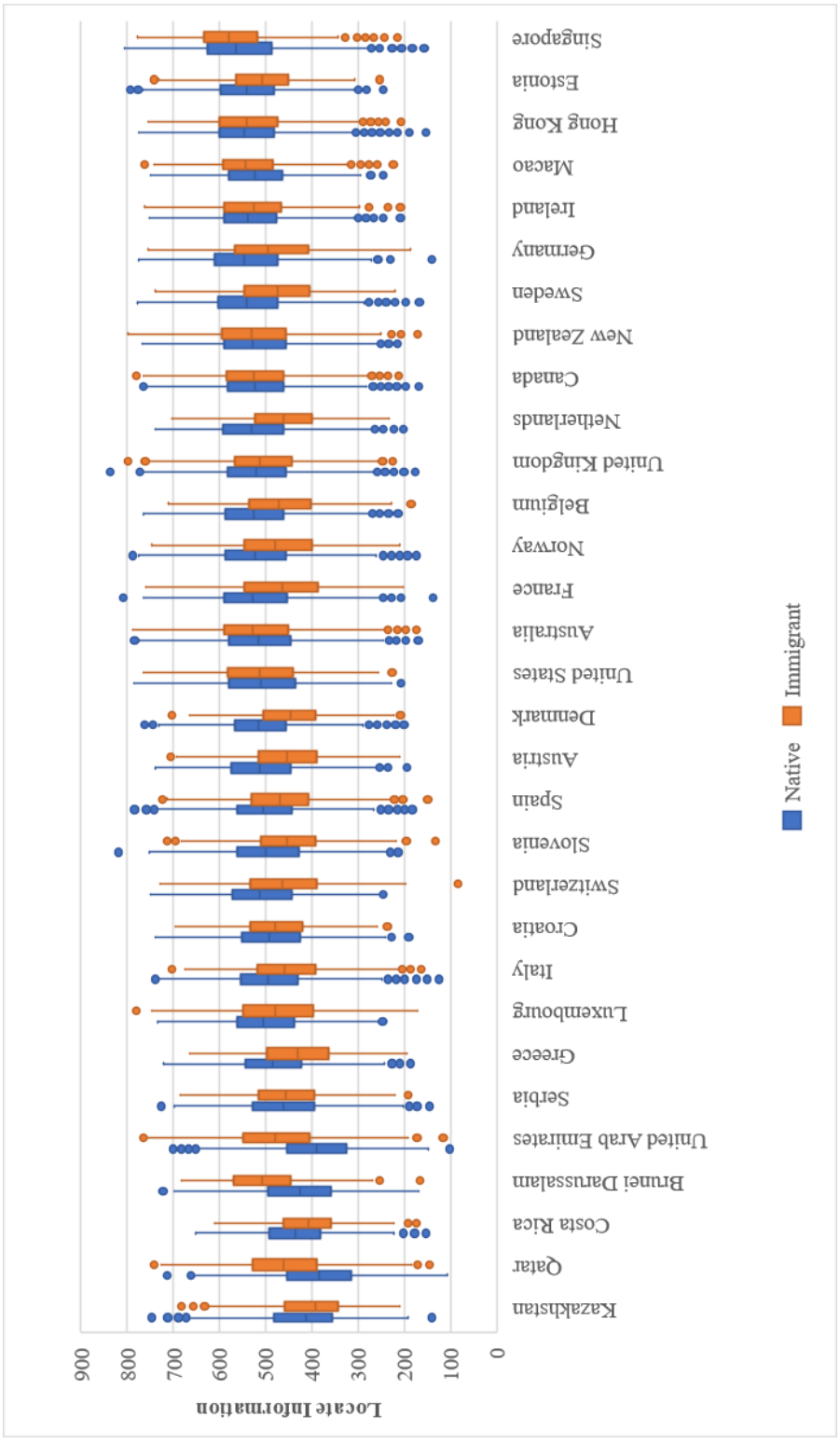
Note. Highlighted values correspond to mean differences larger than 60 units.

According to Table 9, the average values for reading literacy among native students range from 387.6 9 (reading level 1a) to 547.3 (reading level 3) whereas the average values among immigrant students range from 458.9 (level 2) and 565.7 (level 4) suggesting that immigrant students tend to have higher reading literacy than their native peers. Regarding the differences within countries, results show that the largest differences between immigrant and native students are reported in United Arab Emirates (86.2 difference favoring immigrant students), Brunei Darussalam (73.6 difference favoring immigrant students), Qatar (71.3 difference favoring immigrant students), Sweden (67.4 difference favoring native students), Netherlands (66.7 difference favoring native students), and Denmark (63.5 difference favoring native students).

Descriptive statistics were also obtained for each reading subscale. Results for the *locate information* subscale are shown in Figure 17. According to Figure 17, the median values for the reading subscale *locate information* range between 300 and 600. The countries with the highest values include Singapore, Germany, and Hong Kong whereas Costa Rica and Kazakhstan show the lowest. Regarding the immigration status, results show that the countries with the largest difference in median values between immigrant and native students include Qatar, United Arab Emirates, and Brunei Darussalam and the differences are in favor of immigrant students that is, immigrant students within these countries scored higher than their native peers.

Figure 17

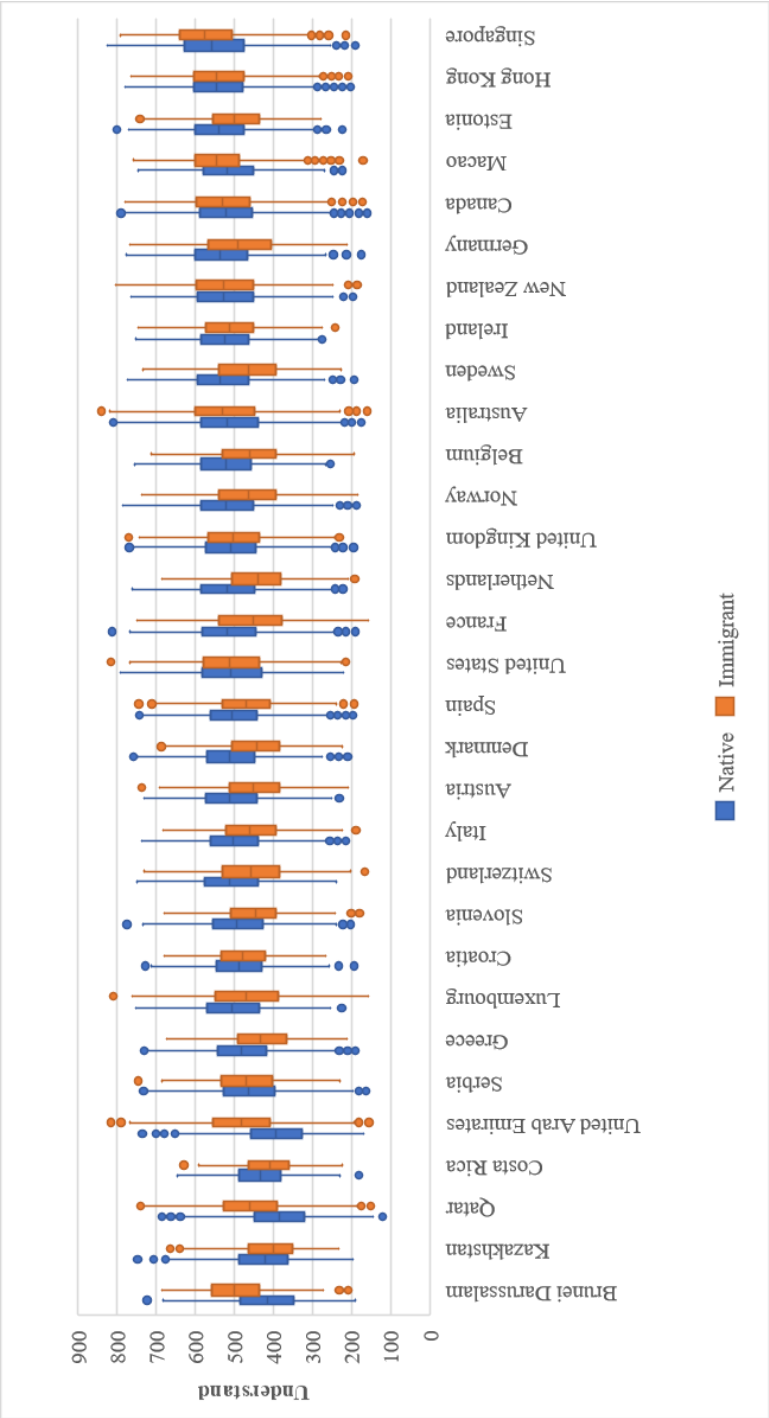
Summary Locate Information Subscale per Country and Immigrant Status



The *understand* subscale was inspected next. Results are shown in Figure 18.

Figure 18

Summary Understand Subscale per Country and Immigrant Status



According to Figure 18, the median values of the *understand* subscale range between 300 and 700 with most values closer to 500. The countries with the lowest median values include Costa Rica and Kazakhstan whereas Canada and Singapore showed the highest median values. Regarding the immigration status, results show that there is variation within countries across native and immigrant students. The largest differences are reported in Brunei Darussalam, Qatar, and United Arab Emirates where immigrant students scored higher than their native peers.

The *evaluate and reflect* subscales were analyzed next. Results are shown in Figure 19.

Figure 19

Summary Evaluate and Reflect Subscale per Country and Immigrant Status

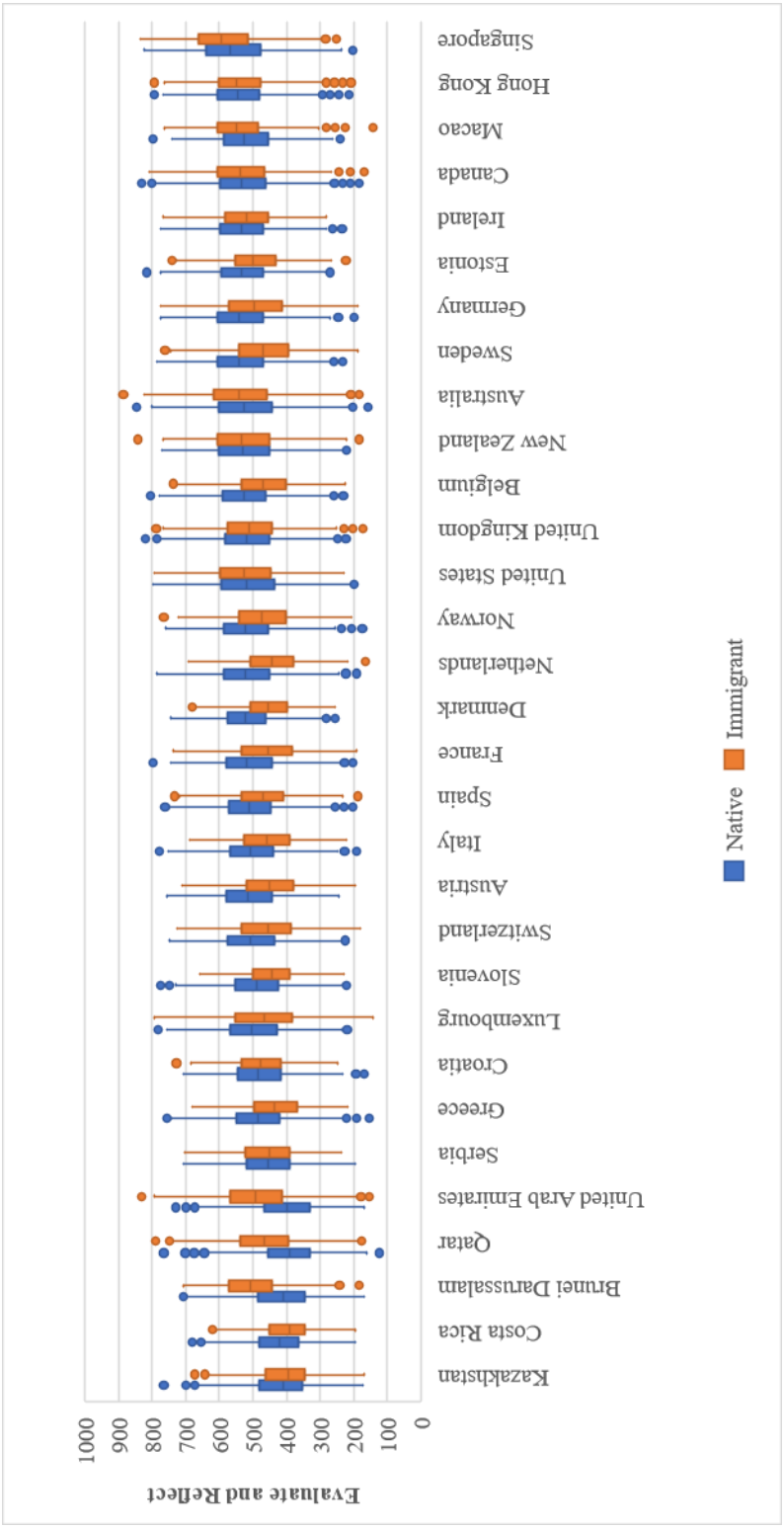


Figure 19 shows that the distribution of scores for the evaluate and reflect subscale follow a consistent pattern with respect to the previous subscales where the median scores are between 300 and 600 across countries and the differences between immigrant and native students remain. The difference continues to be larger in Brunei Darussalam, and in Netherlands, Qatar, and United Arab Emirates. Immigrant students living in these countries show higher median values than their native peers except in Netherlands where the values are higher among native students.

To summarize, results for the reading measure suggest that the distribution of scores follow a similar pattern across countries whereas the distribution changes within countries between immigrant and native students where the former tend to obtain higher scores. Moreover, the differences within countries were consistently larger in Brunei Darussalam, Qatar, and United Arab Emirates. These findings provide preliminary evidence suggesting that the reading measure might not be invariant within countries.

4.2 Evaluation of Measurement Invariance

Measurement invariance was evaluated through two statistical approaches: multiple group confirmatory factor analysis (MGCFA) and the alignment optimization procedure. The analyses were performed in Mplus version 8.1, and all the programming codes are provided in the appendix. Results are described next.

4.2.1 Multiple Group Confirmatory Factor Analysis (MGCFA)

Measurement invariance was initially evaluated through MGCFA following the model trimming strategy that considers different degrees of measurement invariance starting with the evaluation of configural invariance where an initial unconstrained model was evaluated across the countries, followed by the evaluation of metric invariance where

cross-country equality constraints were imposed on the factor loadings, and finally, the evaluation of scalar invariance where the factor loadings and item intercepts (or thresholds in the case of categorical items) were constrained to be equal across countries. The measurement models for the bullying, sense of belonging at school, and reading scales are shown in figures 4, 5, and 6, respectively.

4.2.1.1 Bullying. A single-factor MGCFA model with a mean and threshold structures (see Figure 4) was fitted to evaluate the latent structure of the bullying scale using the weighted least square mean and variance adjusted (WLSMV) estimator and theta parameterization. Measurement invariance of the six ordinal items from the bullying scale was evaluated across 31 countries and the immigration status of the students was not considered given the features of the traditional MGCFA technique.

The analysis started with the evaluation of the configural model, followed by the metric and scalar models. Table 10 shows the contribution to the overall model chi-square by each country. The countries that contributed the most to the chi-square were Italy, Spain, and Canada whereas Macao, Netherlands, and Norway only contributed by 0.84, 1.32, and 1.34%, respectively. Table 11 shows the model fit results for the three models evaluated using a total sample size of 218,315 students.

Table 10

Contribution to Overall Chi-Square per Country- Bullying Scale

Country	Configural Model		Metric Model		Scalar Model	
	X^2 Contribution	%	X^2 Contribution	%	X^2 Contribution	%
Macao	166.16	0.84	353.166	0.75	605.931	1.10
Netherlands	260.74	1.32	857.147	1.81	915.956	1.66
Norway	264.65	1.34	593.184	1.25	627.885	1.14
Switzerland	291.37	1.47	291.536	0.62	337.797	0.61
Greece	305.71	1.54	879.223	1.86	854.026	1.54
Germany	310.04	1.57	303.196	0.64	317.051	0.57
Denmark	378.00	1.91	666.775	1.41	849.804	1.54
Luxembourg	393.08	1.98	438.041	0.93	455.498	0.82
Serbia	397.44	2.01	1265.525	2.67	1582.395	2.86
Brunei Dar.	413.98	2.09	6423.911	13.57	8403.430	15.19
France	420.46	2.12	467.252	0.99	456.177	0.82
Ireland	438.39	2.21	419.785	0.89	497.899	0.90
Estonia	452.47	2.28	281.404	0.59	325.037	0.59
Sweden	462.82	2.34	446.529	0.94	518.053	0.94
Croatia	509.94	2.58	1464.570	3.09	1655.909	2.99
Singapore	522.16	2.64	966.927	2.04	1264.615	2.29
Belgium	526.97	2.66	788.777	1.67	842.298	1.52
US	534.44	2.70	605.394	1.28	693.760	1.25
Austria	547.03	2.76	2393.828	5.06	636.924	1.15
Hong Kong	548.61	2.77	748.866	1.58	1120.283	2.03

Slovenia	594.03	3.00	603.871	1.28	759.655	1.37
New Zealand	648.37	3.27	1598.324	3.38	1879.980	3.40
Costa Rica	811.89	4.10	1233.711	2.61	1513.820	2.74
UK	906.49	4.58	1446.072	3.05	1893.687	3.42
Australia	1053.09	5.32	2393.828	5.06	2695.850	4.87
Kazak.	1053.88	5.32	6234.919	13.17	8754.029	15.83
U. Arab Emirates	1104.69	5.58	1845.039	3.90	2233.145	4.04
Qatar	1116.66	5.64	1842.098	3.89	2162.226	3.91
Italy	1147.94	5.80	1396.512	2.95	1733.161	3.13
Spain	1606.42	8.11	6353.989	13.42	6730.742	12.17
Canada	1615.01	8.16	1735.601	3.67	1999.098	3.61

Note. The three countries with the highest and lowest contribution are highlighted in dark and light grey, respectively.

Table 11

Model Fit for MGCFA Bullying Scale across 31 Countries

Model	X^2	***df***	***p*-value**	**RMSEA**	**RMSEA CI**	**CFI**	**TLI**	**SRMR**
Configural	19802.91	339	0.000	0.090*	[0.089 - 0.091]	0.989	0.985	0.031
Metric	45597.953	354	0.000	0.098*	[0.097 - 0.098]	0.975	0.982	0.055
Metric vs. config.	27767.314	330	0.000	Δ= 0.008	-	Δ= -0.014	-	-
Scalar	55316.120	969	0.000	0.089*	[0.089 - 0.090]	0.969	0.985	0.056
Scalar vs. metric	15246.834	300	0.000	Δ= -0.009	-	Δ= -0.006	-	-

*p<0.001

The configural model was rejected by the chi square test $X^2(339) = 19802.9, p = 0.000$ and the RMSEA (>0.08) indicated poor fit. However, the values of the incremental fit indices (TLI and CFI ≥ 0.90) and the value of the SRMR (<0.05) suggested acceptable fit. Residuals were also inspected to identify possible sources of poor local fit. Reported in Table 12 are the countries for which correlations between the residuals from two items were found to exceed 0.10. According to the results, 13 out of 31 countries showed correlations between residuals that suggest poor local fit specifically, Costa Rica, Germany and the Netherlands showed the highest number of residual correlations exceeding 0.10.

Moreover, the correlation residual between items 1 and 5 (“left out” and “got hit/pushed”, respectively) was flagged in 11 out of the 13 countries that showed problems related to correlation residuals and the correlation was negative suggesting that the proposed model overpredicts the sample polychoric correlation between these items by the values reported in Table 12. This finding also suggests that these items might be sharing a common variable that was not considered in the model in those countries only. The residual correlation between items 1 and 4 (“left out” and “took/destroyed things”, respectively) was also found in four of the 13 countries reported whereas the residual correlation between items 1 and 2 (“left out” and “students made fun of me”, respectively) was reported for two countries. Item 1 (“left out”) is reported for all the countries suggesting that this item might be the most problematic with respect to the overall model fit.

Table 12

Summary Inspection of Residuals- Configural, Metric, and Scalar Models- Bullying Scale

Country	Configural Model		Metric Model		Scalar Model	
	Count	**Correlation Residuals > .10**	**Count**	**Correlation Residuals > .10**	**Count**	**Correlation Residuals > .10**
Brunei D.	-	-	-	-	15*	$X_1 - X_5 = -0.342$ $X_2 - X_5 = -0.309$ $X_1 - X_3 = -0.306$
Canada	-	-	1	$X_1 - X_5 = -0.132$	1	$X_1 - X_5 = -0.135$
Croatia	-	-	1	$X_1 - X_2 = 0.148$	2	$X_1 - X_2 = 0.168$ $X_1 - X_6 = 0.116$
Hong Kong	-	-	1	$X_1 - X_6 = 0.159$	2	$X_1 - X_6 = 0.135$ $X_2 - X_3 = -0.113$
Luxem.	-	-	1	$X_1 - X_2 = 0.116$	1	$X_1 - X_2 = 0.128$
Macao	-	-	-	-	1	$X_2 - X_5 = -0.108$
Singapore	-	-	1	$X_1 - X_5 = -0.103$	1	$X_1 - X_5 = -0.111$
Switzer.	-	-	-	-	1	$X_1 - X_2 = 0.106$
U. Arab.	-	-	-	-	1	$X_1 - X_5 = -0.108$
UK	-	-	1	$X_1 - X_2 = -0.137$	1	$X_1 - X_5 = -0.148$
Australia	-	-	2	$X_1 - X_5 = -0.156$ $X_6 - X_5 = -0.112$	2	$X_1 - X_5 = -0.164$ $X_6 - X_5 = -0.124$
Denmark	-	-	2	$X_1 - X_4 = -0.116$ $X_1 - X_5 = -0.125$	4*	$X_1 - X_4 = -0.134$ $X_1 - X_5 = -0.142$ $X_6 - X_5 = -0.113$
Greece	-	-	-	-	2	$X_2 - X_3 = -0.104$ $X_2 - X_4 = -0.106$
US	-	-	2	$X_1 - X_2 = 0.117$ $X_5 - X_6 = -0.103$	3	$X_1 - X_2 = 0.117$ $X_1 - X_5 = -0.101$ $X_6 - X_5 = -0.108$
New Z.	-	-	4*	$X_1 - X_4 = -0.109$ $X_1 - X_5 = -0.160$	6*	$X_1 - X_4 = -0.123$ $X_{\quad} - X_{\quad} = -0.174$

				$X_5 - X_6 = -0.126$		$X_6 - X_5 = -0.142$
Kazak.	-	-	5*	$X_1 - X_3 = -0.186$ $X_1 - X_4 = -0.186$ $X_1 - X_5 = -0.191$	4*	$X_1 - X_3 = -0.160$ $X_1 - X_4 = -0.156$ $X_1 - X_5 = -0.163$
Austria	1	$X_1 - X_5 = -0.103$	2	$X_1 - X_2 = 0.188$ $X_6 - X_4 = -0.132$	1	$X_1 - X_2 = 0.127$
France	1	$X_1 - X_4 = -0.116$	1	$X_1 - X_2 = 0.141$	1	$X_1 - X_2 = 0.146$
Ireland	1	$X_1 - X_5 = -0.103$	2	$X_1 - X_4 = -0.138$ $X_1 - X_5 = -0.134$	2	$X_1 - X_4 = -0.141$ $X_1 - X_5 = -0.145$
Norway	1	$X_1 - X_5 = -0.121$	2	$X_1 - X_2 = 0.123$ $X_1 - X_6 = 0.141$	2	$X_1 - X_2 = 0.123$ $X_1 - X_6 = 0.141$
Qatar	1	$X_1 - X_2 = 0.105$	1	$X_1 - X_5 = -0.108$	2	$X_1 - X_5 = -0.117$ $X_2 - X_5 = -0.103$
Spain	1	$X_1 - X_5 = -0.101$	2	$X_1 - X_2 = 0.181$ $X_1 - X_6 = 0.120$	2	$X_1 - X_2 = 0.194$ $X_1 - X_6 = 0.134$
Belgium	2	$X_1 - X_4 = -0.112$ $X_1 - X_5 = -0.134$	10*	$X_1 - X_3 = -0.227$ $X_1 - X_5 = -0.259$ $X_2 - X_5 = -0.231$	3	$X_1 - X_2 = 0.181$ $X_6 - X_4 = -0.145$ $X_6 - X_5 = -0.107$
Estonia	2	$X_1 - X_4 = -0.105$ $X_1 - X_5 = -0.103$	1	$X_1 - X_5 = -0.102$	1	$X_1 - X_5 = -0.107$
Italy	2	$X_1 - X_2 = 0.103$ $X_1 - X_5 = -0.110$	2	$X_1 - X_4 = -0.115$ $X_1 - X_5 = -0.101$	2	$X_1 - X_2 = 0.111$ $X_1 - X_4 = -0.102$
Sweden	2	$X_1 - X_4 = -0.102$ $X_1 - X_5 = -0.162$	2	$X_1 - X_5 = -0.151$ $X_1 - X_6 = 0.125$	2	$X_1 - X_5 = -0.159$ $X_6 - X_1 = 0.117$
Costa Rica	3	$X_1 - X_4 = -0.138$ $X_1 - X_5 = -0.125$ $X_2 - X_4 = -0.106$	2	$X_1 - X_2 = 0.158$ $X_4 - X_6 = -0.103$	2	$X_1 - X_2 = 0.160$ $X_6 - X_4 = -0.105$
Germany	3	$X_1 - X_4 = -0.144$ $X_1 - X_5 = -0.139$ $X_2 - X_5 = -0.105$	1	$X_1 - X_2 = 0.151$	2	$X_1 - X_2 = 0.149$ $X_6 - X_4 = -0.104$
Nether.	3	$X_1 - X_5 = -0.130$ $X_4 - X_6 = -0.102$ $X_5 - X_6 = -0.121$	3	$X_1 - X_2 = 0.184$ $X_1 - X_6 = 0.203$ $X_5 - X_6 = -0.109$	3	$X_1 - X_2 = 0.185$ $X_1 - X_6 = 0.205$ $X_6 - X_5 = -0.113$

* The three largest correlations are reported.

The metric model was evaluated next. The factor loadings were constrained to equality across groups and the results are reported in Table 11. The "difftest" option in Mplus was used to test this model and compare it against the configural model, this option generates correct values of the chi-square difference statistic.

The metric model was also rejected by the chi-square test $X^2 = 45597.953\ (354), p = 0.000$, and the largest contribution to the overall chi-square was from Brunei Darussalam, Spain, and Kazakhstan, whereas Estonia, Switzerland, and Germany showed the smallest contribution.

The fit of the metric model was not statistically worse than the fit of the less restrictive configural model based on the change in RMSEA ($\Delta RMSEA = 0.008, \leq 0.05$) however, the change in CFI ($\Delta CFI = -0.01, \sim \geq -0.004$) and the value of the X^2 suggested worse fit moreover, the inspection of residuals (see Table 12) showed indications of severe misspecification for 24 out of 31 countries where Belgium and Kazakhstan reported the highest number of correlations between residuals (10 and 5, respectively) higher than 0.1.

It is interesting that six out of the 10 the correlations between residuals reported for Belgium exceeded 0.2 and all were negative suggesting the metric model might be overpredicting the polychoric correlations between those pair of items by the values shown in Table 12. A total of 11 countries that had not reported indications of misspecification based on the residual correlations in the configural model, were flagged in the metric model. The findings from the residuals suggest that the pattern coefficients might not be equivalent among these countries that is, the target latent construct is manifested in different ways in each of these countries.

Finally, the scalar model was tested. To do so, equality constraints were also imposed on the unstandardized thresholds that had not been previously restricted and as with the previous model, the "difftest" option was also used to compare the fit of this model to that of the metric model. This model was also rejected by the chi-square test, $X^2(969) = 55316.120, p = 0.000$ (see Table 11). The largest contribution to the overall chi-square was by Kazakhstan, Brunei Darussalam, and Spain while Germany, Estonia, and Switzerland showed the lowest contribution (see Table 10). The model fit when compared to the metric model was not statistically worse based on the change in RMSEA ($\Delta RMSEA = -0.009, \leq 0.01$). However, the change in CFI suggested worse fit based on the cutoff value for the scalar model: $\Delta CFI = -0.006, \sim \geq -0.004$ (see Table 11).

Regarding the correlations between residuals in the scalar model, Table 12 shows that 29 out of the 31 countries showed residual correlations larger than 0.1 and all the residual correlations reported for Belgium were flagged suggesting severe local fit problems in this country; Switzerland and United Arab Emirates on the other hand, reported the smallest number of problems related to local fit. The overall results suggest that the scalar model tends to both over and underpredict the polychoric correlations among the items and the empirical evidence is not sufficient to retain the model therefore, the scalar invariance hypothesis should be rejected. These findings indicate that it is likely that students from different countries might be using the response scale of the items in a different way thus, the estimated factor means are in turn, likely to be biased.

4.2.1.2 Sense of Belonging at School. As with the exposure to bullying scale, a single-factor MGCFA with a mean and threshold structures (Figure 5) was fitted to the data from the sense of belonging at school scale using the weighted least square mean

and variance adjusted (WLSMV) estimator and theta parameterization given the scale of the item responses. Measurement invariance of the scale was evaluated across the 31 countries following the model trimming strategy where a configural model was initially fitted with no equality constraints followed by increasingly constrained models: metric and scalar.

The contribution to the overall model chi-square per country is shown in Table 13. According to the results, Qatar, Kazakhstan, and United Arab Emirates made the largest contribution to the chi-square test whereas Germany, Switzerland, and Denmark contributed the least. The values of the fit statistics for the configural model are shown in Table 14.

Table 13

Contribution to Overall Chi-Square per Country- Sense of Belonging at School Scale

Country	Configural Model		Metric Model	
	X^2 Contribution	%	X^2 Contribution	%
Germany	506.07	0.43	812.460	0.62
Switzerland	836.12	0.71	992.056	0.75
Denmark	1050.94	0.89	2627.053	2.00
Macao	1121.49	0.95	1406.678	1.07
Ireland	1178.67	1.00	1291.405	0.98
Netherlands	1190.05	1.01	2161.807	1.65
Austria	1259.48	1.07	2092.025	1.59
New Zealand	1353.80	1.14	1378.241	1.05
Belgium	1507.41	1.27	1895.360	1.44
Brunei Dar.	1564.78	1.32	3723.888	2.83
France	1766.88	1.49	4521.855	3.44
Sweden	1843.55	1.56	1994.183	1.52
Norway	1956.54	1.65	3056.892	2.33
Luxembourg	2015.10	1.70	1796.122	1.37
Estonia	2049.37	1.73	1705.456	1.30
US	2090.63	1.77	2196.373	1.67
Italy	2091.30	1.77	2597.645	1.98
Greece	2166.76	1.83	2463.865	1.88
Costa Rica	2751.64	2.33	2416.310	1.84
Croatia	2845.23	2.41	2809.183	2.14
Slovenia	3008.19	2.54	2479.765	1.89
Singapore	3188.14	2.70	2630.019	2.00
Hong Kong	3563.35	3.01	4194.097	3.19
UK	3653.88	3.09	3983.931	3.03
Australia	5151.61	4.36	4748.095	3.61
Serbia	5685.32	4.81	4259.379	3.24
Canada	6871.60	5.81	6039.044	4.60
Spain	7387.64	6.25	19639.246	14.95
Qatar	13478.10	11.40	11457.744	8.72
Kazak.	15909.10	13.45	12348.421	9.40
U. Arab. Emirates	17217.80	14.56	15682.510	11.93

Note. The three countries with the highest and lowest contribution are highlighted in dark and light grey, respectively.

Table 14

Model Fit for MGCFA Sense of Belonging at School Scale across 31 Countries

Model	X^2	***df***	***p*-value**	**RMSEA**	**RMSEA CI**	**CFI**	**TLI**	**SRMR**
Configural	118260.537	339	0.000	0.222*	[0.221 – 0.223]	0.908	0.874	0.075
Metric	1285352.93	465	0.000	0.167*	[0.166 – 0.1167]	0.898	0.929	0.084
Metric vs. config.	38701.147	330	0.000	$\Delta = -0.055$	-	$\Delta = -0.01$	-	-

*p<0.001

Results show that the configural model is rejected by the chi-square test ($X^2(339) = 118260.53, p = 0.000$). Regarding the incremental fit indexes, results show that the CFI supports the configural model ($CFI \geq 0.90$) whereas the estimated values for the TLI and SRMR do not suggest acceptable fit ($TLI < 0.90;\ SRMR > 0.05$). The correlations between residuals were also inspected. Results are shown in Table 15.

Table 15

Summary Inspection of Residuals- Sense of Belonging at School Scale

	Configural Model		Metric Model	
Country	**Count**	**Correlation Residuals > .10**	**Count**	**Correlation Residuals > .10**
Austria	2	$X_2 - X_3 = -0.120$ $X_2 - X_4 = -0.139$	2	$X_2 - X_4 = -0.157$ $X_2 - X_5 = 0.125$
Denmark	2	$X_2 - X_5 = 0.107$ $X_3 - X_6 = -0.121$	3	$X_1 - X_6 = 0.125$ $X_5 - X_4 = -0.115$ $X_5 - X_6 = -0.113$
Switzer.	3	$X_2 - X_4 = -0.140$ $X_2 - X_5 = 0.140$ $X_4 - X_5 = -0.129$	6*	$X_2 - X_4 = -0.208$ $X_4 - X_5 = -0.121$ $X_1 - X_2 = -0.118$
Belgium	3	$X_2 - X_4 = -0.123$ $X_2 - X_5 = 0.154$ $X_4 - X_5 = -0.127$	4*	$X_2 - X_5 = 0.190$ $X_1 - X_3 = -0.127$ $X_3 - X_6 = -0.127$
Germany	4*	$X_1 - X_3 = -0.162$ $X_2 - X_4 = -0.153$ $X_2 - X_5 = 0.133$	4*	$X_2 - X_4 = -0.217$ $X_1 - X_3 = -0.130$ $X_4 - X_5 = -0.119$
Ireland	4*	$X_2 - X_5 = 0.145$ $X_5 - X_6 = -0.106$ $X_2 - X_6 = -0.105$	2	$X_2 - X_5 = 0.187$ $X_3 - X_6 = -0.107$
Macao	4*	$X_2 - X_5 = 0.136$ $X_4 - X_5 = -0.132$ $X_1 - X_4 = 0.106$	5*	$X_4 - X_5 = -0.169$ $X_2 - X_5 = 0.146$ $X_1 - X_3 = -0.130$
New Z.	4*	$X_2 - X_5 = 0.136$ $X_4 - X_5 = -0.121$ $X_5 - X_6 = -0.115$	2	$X_2 - X_5 = 0.158$ $X_3 - X_6 = -0.103$
Spain	4*	$X_2 - X_4 = -0.127$ $X_1 - X_2 = -0.109$ $X_2 - X_6 = -0.105$	6*	$X_2 - X_4 = -0.159$ $X_2 - X_3 = 0.150$ $X_4 - X_5 = -0.135$
Sweden	4*	$X_2 - X_4 = -0.125$ $X_1 - X_5 = -0.121$ $X_1 - X_3 = -0.110$	4*	$X_2 - X_5 = 0.174$ $X_1 - X_3 = -0.143$ $X_3 - X_5 = 0.124$
Canada	5*	$X_5 - X_6 = -0.128$ $X_4 - X_5 = -0.128$ $X_2 - X_5 = 0.133$	6*	$X_2 - X_5 = 0.159$ $X_4 - X_5 = -0.112$ $X_2 - X_4 = -0.108$
Italy	5*	$X_2 - X_5 = 0.135$ $X_2 - X_4 = -0.122$ $X_3 - X_6 = -0.119$	6*	$X_2 - X_4 = -0.189$ $X_4 - X_5 = -0.144$ $X_1 - X_6 = 0.134$
UK	5*	$X_2 - X_5 = 0.134$ $X_4 - X_5 = -0.119$ $X_2 - X_6 = -0.102$	2	$X_2 - X_5 = 0.200$ $X_3 - X_6 = -0.146$
Nether.	6*	$X_2 - X_5 = 0.171$ $X_6 - X_3 = -0.153$ $X_2 - X_4 = -0.128$	6*	$X_3 - X_6 = -0.176$ $X_1 - X_6 = 0.169$ $X_1 - X_3 = -0.162$
Singapore	6*	$X_4 - X_5 = -0.162$ $X_2 - X_5 = 0.153$ $X_2 - X_4 = -0.140$	7*	$X_2 - X_5 = 0.168$ $X_4 - X_5 = -0.162$ $X_2 - X_4 = -0.139$
US	6*	$X_4 - X_5 = -0.155$	7*	$X_2 - X_5 = 0.184$

		$X_5 - X_6 = -0.132$		$X_5 - X_6 = -0.122$
Australia	7*	$X_4 - X_5 = -0.150$ $X_2 - X_6 = -0.143$ $X_5 - X_6 = -0.138$	8*	$X_2 - X_5 = 0.186$ $X_2 - X_6 = -0.149$ $X_1 - X_2 = -0.122$
France	7*	$X_1 - X_2 = -0.249$ $X_4 - X_5 = -0.149$ $X_2 - X_5 = 0.147$	6*	$X_1 - X_2 = -0.356$ $X_1 - X_5 = -0.266$ $X_2 - X_5 = 0.251$
Greece	7*	$X_1 - X_2 = -0.204$ $X_2 - X_4 = -0.156$ $X_4 - X_5 = -0.133$	9*	$X_1 - X_2 = -0.256$ $X_2 - X_3 = 0.208$ $X_1 - X_5 = -0.178$
Costa Rica	8*	$X_2 - X_4 = -0.152$ $X_4 - X_5 = -0.137$ $X_2 - X_5 = 0.132$	7*	$X_1 - X_5 = -0.193$ $X_5 - X_6 = -0.170$ $X_2 - X_3 = 0.169$
Hong Kong	8*	$X_4 - X_5 = -0.202$ $X_2 - X_4 = -0.190$ $X_2 - X_5 = 0.174$	8*	$X_2 - X_5 = 0.236$ $X_3 - X_6 = -0.203$ $X_4 - X_5 = -0.184$
Norway	8*	$X_1 - X_2 = -0.142$ $X_1 - X_5 = -0.121$ $X_4 - X_5 = -0.119$	6*	$X_1 - X_5 = -0.169$ $X_4 - X_5 = -0.143$ $X_5 - X_6 = -0.142$
Croatia	9*	$X_4 - X_5 = -0.172$ $X_2 - X_4 = -0.170$ $X_2 - X_5 = 0.151$	9*	$X_4 - X_5 = -0.176$ $X_5 - X_6 = -0.156$ $X_1 - X_5 = -0.150$
Estonia	9*	$X_3 - X_5 = 0.142$ $X_6 - X_5 = -0.123$ $X_4 - X_5 = -0.118$	6*	$X_3 - X_5 = 0.176$ $X_2 - X_4 = -0.131$ $X_1 - X_2 = -0.125$
Brunei D.	10*	$X_1 - X_2 = -0.166$ $X_2 - X_5 = 0.143$ $X_2 - X_3 = 0.132$	10*	$X_1 - X_2 = -0.179$ $X_3 - X_4 = -0.178$ $X_1 - X_3 = -0.169$
Luxem.	10*	$X_2 - X_5 = -0.183$ $X_2 - X_4 = -0.168$ $X_4 - X_5 = -0.167$	8*	$X_2 - X_4 = -0.224$ $X_4 - X_5 = -0.178$ $X_2 - X_5 = 0.162$
Slovenia	11*	$X_1 - X_3 = -0.169$ $X_2 - X_5 = 0.166$ $X_3 - X_5 = 0.162$	12*	$X_1 - X_3 = -0.229$ $X_3 - X_6 = -0.221$ $X_2 - X_5 = 0.169$
Kazak.	12*	$X_1 - X_3 = -0.208$ $X_3 - X_4 = -0.206$ $X_4 - X_5 = -0.212$	14*	$X_5 - X_6 = -0.232$ $X_1 - X_3 = -0.223$ $X_1 - X_5 = -0.220$
Qatar	13*	$X_1 - X_2 = -0.262$ $X_2 - X_4 = -0.255$ $X_4 - X_5 = -0.243$	14*	$X_2 - X_4 = -0.279$ $X_1 - X_2 = -0.274$ $X_4 - X_5 = -0.269$
Serbia	13*	$X_2 - X_4 = -0.264$ $X_4 - X_5 = -0.245$ $X_1 - X_2 = -0.224$	14*	$X_5 - X_6 = -0.259$ $X_3 - X_6 = -0.239$ $X_2 - X_6 = -0.238$
U. Arab.	13*	$X_1 - X_2 = -0.245$ $X_1 - X_5 = -0.239$ $X_2 - X_4 = -0.238$	12*	$X_4 - X_5 = -0.331$ $X_2 - X_4 = -0.300$ $X_5 - X_6 = -0.278$

* The three largest correlations are reported.

According to the results all the countries had correlations between pairs of item residuals >.10. Qatar, Serbia, and United Arab Emirates showed the highest number of flagged residual correlations (13 each) whereas Austria and Denmark showed the lowest

number. Given that the flagged correlations were both positive and negative, the model could be under and overpredicting the sample polychoric correlations by the amounts shown in Table 15 which in turn suggests that the flagged correlations probably share an omitted cause for the flagged pairs of items within the countries.

The residual for item 2 ("make friends easily") repeatedly appears across most countries indicating that this item could be particularly problematic in terms of the model fit. The overall inspection of residuals suggests that the proposed model might not be accurately predicting the univariate proportions of item responses. The metric model was fitted next and as with the previous scale, the "difftest" option in Mplus was specified to generate correct values of the scaled chi-square difference statistics. The contribution of each country to the chi-square test is reported in Table 13.

Spain, United Arab Emirates, and Kazakhstan made the largest contributions to the chi-square test whereas Germany, Switzerland, and Ireland made the smallest. The model fit statistics are reported in Table 14. The unstandardized factor loadings for the six items were constrained to equality across countries in the metric model, which was rejected by the chi-square test, $X^2(330) = 38701.147, p = 0.000$. Regarding the incremental fit indexes, results show that $\Delta CFI = -0.01$ is outside the cutoff point ($\Delta CFI \geq -0.004$) and does not support the metric model.

However, the change in RMSEA ($\Delta RMSEA = -0.055$) provides evidence of improvement in model fit based on the suggested cutoff value ($\Delta RMSEA \leq 0.05$). The residual correlations were also inspected to identify possible sources of severe misspecification; results are provided in Table 15. According to the results, all the countries showed residual correlations larger than 0.1, the three countries with the largest

number of problematic residual correlations were Kazakhstan, Qatar, and Serbia whereas the three countries with the smallest number were Austria, Ireland, and New Zealand. Given that (a) the inspection of residuals suggests indications of severe misspecification across most countries, (b) the residual correlations indicate that the metric model could be both under and overestimating the polychoric correlations, and (c) the overall fit is not consistently suggesting a significant improvement in model fit, the metric model was not retained thus, the evaluation of the next model (scalar) was not performed.

4.2.1.3 Reading Literacy. A single-factor MGCFA with a mean and threshold structures and three indicators (Figure 6) was fitted to the data from the reading literacy scale across the 31 countries. The model trimming strategy was followed to evaluate the measurement invariance of the scale with increasingly restrictive models. Given that the data of this scale were continuous, the maximum likelihood (ML) estimator was used. The model with 31 countries did not converge; thus, one country was removed at a time until convergence was achieved. The maximum number of countries required for convergence was 12 therefore, the analyses were conducted for the following countries: Australia, Austria, Belgium, Brunei Darussalam, Canada, Costa Rica, Croatia, Denmark, Estonia, France, Germany, and Greece.

As with the previous analysis, the analysis of measurement invariance began with the evaluation of a configural model, followed by metric and scalar models. Table 16 shows the contribution to the overall model chi-square by each country.

Table 16

Contribution to Overall Chi-Square per Country- Reading Literacy Scale

Country	Configural Model		Metric Model		Scalar Model	
	X^2 Contribution	%	X^2 Contribution	%	X^2 Contribution	%
Austria	0.000	0.00	33.738	0.74	105.084	0.91
Greece	0.000	0.00	46.869	1.02	132.471	1.15
Estonia	0.000	0.00	88.690	1.93	880.325	7.65
Belgium	0.000	0.00	106.862	2.33	364.509	3.17
Germany	0.000	0.00	216.357	4.72	332.800	2.89
Brunei Dar.	0.000	0.00	273.975	5.98	704.860	6.12
France	0.000	0.00	344.033	7.50	725.187	6.30
Costa Rica	0.000	0.00	355.604	7.76	2445.981	21.25
Denmark	0.000	0.00	468.738	10.23	711.062	6.18
Australia	0.000	0.00	621.632	13.56	1957.084	17.00
Canada	0.000	0.00	985.191	21.49	1353.424	11.76
Croatia	0.000	0.00	1042.394	22.74	1799.014	15.63

Note. The three countries with the highest and lowest contribution are highlighted in dark and light grey, respectively.

The configural model for the three items from the reading literacy scale was just identified. Thus, none of the countries contributed to the overall Chi-square. The values of the fit statistics for the configural model are shown in Table 17.

Table 17

Model Fit for MGCFA Reading Literacy Scale across 12 Countries

Model	X^2	***df***	***p*-value**	**RMSEA**	**RMSEA CI**	**CFI**	**TLI**	**SRMR**
Configural	0.000	0	0.000	0.000	[0.000-0.000]	1.0	1.0	0.000
Metric	4584.082	22	0.000	0.159	[0.155 – 0.163]	0.991	0.986	0.087
Metric vs. config.	4584.082	22	0.000	-	-	Δ= -0.009	-	-
Scalar	11512.112	44	0.000	0.178	[0.175 – 0.181]	0.979	0.982	0.089
Scalar vs. metric	6928.03	22	0.000	Δ= 0.019	-	Δ= -0.012	-	-

The fit statistics for the configural model confirmed that the model is just identified given that the model only includes three indicators. The evaluation of invariance proceeded, and equality constraints were imposed on the factor loadings to test the metric model across the countries. The contribution of each country to the overall Chi-square is shown in Table 16; the highest contribution was made by Australia, Canada, and Croatia whereas Austria, Greece, and Estonia showed the lowest. The fit statistics from the metric model are shown in Table 17. Results show that the model was rejected by the Chi-square test $X^2(22) = 4584.082, p = 0.000$. Regarding the change in CFI ($\Delta CFI \leq -0.004$) results also suggested poor fit that is, the relative fit of the metric model was statistically worse than that of the configural model. Residuals were inspected next, and results are reported in Table 18.

Table 18

Summary Inspection of Residuals- Configural, Metric, and Scalar Models- Reading Scale

Country	Configural Model		Metric Model		Scalar Model	
	Count	**Correlation Residuals z-score > 1.96**	**Count**	**Correlation Residuals z-score > 1.96**	**Count**	**Correlation Residuals z-score > 1.96**
Australia	0	0	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 8.811$	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 8.216$
Austria	0	0	2	$X_1 - X_3 = 4.040$ $X_1 - X_2 = 3.231$	2	$X_1 - X_2 = 3.687$ $X_1 - X_3 = 3.568$
Belgium	0	0	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 2.404$	2	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$
Brunei D.	0	0	3	$X_1 - X_2 = 7.153$ $X_1 - X_3 = 5.233$ $X_2 - X_3 = 999.0$	3	$X_1 - X_2 = 7.384$ $X_1 - X_3 = 4.818$ $X_2 - X_3 = 999.0$
Canada	0	0	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 11.244$	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 10.715$
Costa Rica	0	0	3	$X_1 - X_2 = 7.653$ $X_1 - X_3 = 11.094$ $X_2 - X_3 = 6.623$	3	$X_1 - X_2 = 7.749$ $X_1 - X_3 = 10.653$ $X_2 - X_3 = 5.748$
Croatia	0	0	3	$X_1 - X_2 = 11.015$ $X_1 - X_3 = 11.834$ $X_2 - X_3 = -1.707$	3	$X_1 - X_2 = 11.034$ $X_1 - X_3 = 11.494$ $X_2 - X_3 = 999.0$
Denmark	0	0	3	$X_1 - X_2 = 5.165$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 999.0$	3	$X_1 - X_2 = 5.752$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 999.0$
Estonia	0	0	2	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$	3	$X_1 - X_2 = 999.0$ $X_1 - X_3 = 999.0$ $X_2 - X_3 = 999.0$
France	0	0	3	$X_1 - X_2 = 7.753$ $X_1 - X_3 = 5.317$ $X_2 - X_3 = 999.0$	3	$X_1 - X_2 = 7.890$ $X_1 - X_3 = 4.849$ $X_2 - X_3 = 999.0$
Germany	0	0	3	$X_1 - X_2 = 6.428$ $X_1 - X_3 = 4.895$ $X_2 - X_3 = 999.0$	3	$X_1 - X_2 = 6.674$ $X_1 - X_3 = 4.644$ $X_2 - X_3 = 999.0$
Greece	0	0	3	$X_1 - X_2 = 4.174$ $X_1 - X_3 = 3.016$ $X_2 - X_3 = 999.0$	3	$X_1 - X_2 = 4.529$ $X_1 - X_3 = 2.469$ $X_2 - X_3 = 999.0$

All the countries under analysis showed statistically significant standardized residuals based on the z-scores for the pairs of indicators shown in Table 18 and most of

the correlations between residuals were positive suggesting that the metric model might be underpredicting the observed correlations.

The scalar model was analyzed next where the intercepts of each indicator were constrained to equality across the groups. The contribution of each country to the overall Chi-square is shown in Table 16. Results show that Costa Rica, Australia, and Croatia were the countries that contributed the most to the overall Chi-square while Austria, Greece, and Germany contributed the least. Results related to the model fit of the scalar model are shown in table 17. The scalar model was rejected by the Chi-square $X^2(44) = 11512.112, p = 0.000$. Regarding the change in RMSEA, results show that the fit of the scalar model improved over the metric model ($\Delta RMSEA = 0.019$) however the change in CFI ($\Delta CFI = -0.012$) did not suggest model fit improvement.

The residuals were also inspected, results are shown in Table 18. Results indicate that there are local fit problems, and the scalar model is mostly underpredicting the observed correlations. The overall results do not provide evidence of measurement invariance suggesting that the indicators might not have the same meaning across countries therefore, test score comparisons among countries are likely to be biased and the cultural differences are likely to impact test performance. Direct comparisons among countries are likely to lead to biased interpretations of test results.

4.2.2 The Alignment Optimization

The alignment optimization was implemented following the procedures reported by Munck, Barber, and Torney-Purta (2018); Lomazzi (2018); and Roberson and Zumbo (2019) who evaluated measurement invariance of international measures across countries and within specific subgroups of students within the countries (e.g., male/female,

migration status, cohort). To do so, they generated a grouping variable to reflect the comparative design of their studies where students were to be compared based on the country and the other relevant variables (i.e., sex, cohort).

Given the purpose of this dissertation, the grouping variable was generated so that it would consider the country and the immigration status (native or immigrant). The number of groups to be included in the alignment analysis were $n(countries) \times n(immigration\ status)$ that is, 31 $countries\ \times 2\ levels\ of\ immigration\ status$. A total of 62 groups were generated and included in the alignment optimization procedure.

4.2.2.1 Bullying. Measurement invariance for the six items from the bullying scale was evaluated through alignment optimization. As previously mentioned, a total of ***62 groups*** were included in the analysis to account for the country and immigration status thus, the model that was evaluated in this analysis only included the latent factor (bullying) and its six indicators as shown in Figure 7. All the analyses were conducted in Mplus version 8.1. The input file with the code for the bullying scale is provided in appendix C.

The procedure began with a configural model where the factor means and variances for each group were fixed to zero and one, respectively while the factor loadings and item thresholds were freely estimated. Then, the alignment procedure continued by freeing the factor means and variances and selecting the values that minimized the total amount of noninvariance through the simplicity function. The estimator used in this analysis was the robust maximum likelihood (MLR) and free alignment was used initially to identify the baseline group that would prevent misspecification of the model. Results from the fixed alignment suggested that the group

of immigrant students living in Netherlands (labeled as 52800) should be set as the baseline group to identify the model thus, the fixed alignment was implemented following this suggestion.

Table 19 shows the groups for which the factor loadings were ***invariant*** according to the alignment optimization procedure.

Table 19

Summary Invariant Factor Loadings Bullying Scale

Country	Item 6		Item 3		Item 1		Item 5		Item 2		Item 4	
	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.
Belgium	x	x	x	x	x	x	x	x	x	x	x	x
Canada	x	x	x	x	x	x	x	x	x	x	x	x
Costa Rica	x	x	x	x	x	x	x	x	x	x	x	x
France	x	x	x	x	x	x	x	x	x	x	x	x
Ireland	x	x	x	x	x	x	x	x	x	x	x	x
New Zealand	x	x	x	x	x	x	x	x	x	x	x	x
United King.	x	x	x	x	x	x	x	x	x	x	x	x
United States	x	x	x	x	x	x	x	x	x	x	x	x
Denmark	x	x		x	x	x	x	x	x	x	x	x
Norway	x	x	x	x	x	x	x	x	x	x		x
Singapore	x	x	x	x	x	x	x	x	x	x		x
Austria	x	x	x	x	x	x	x	x		x		x
Croatia	x	x	x	x	x	x		x	x	x		x
Germany	x	x	x	x	x	x	x	x	x	x		
Luxembourg	x	x	x	x	x	x	x	x	x		x	
Netherlands	x	x		x	x	x	x	x	x	x	x	
Australia	x	x			x	x	x	x	x	x	x	
Estonia	x	x	x	x	x	x		x	x	x		
Macao	x	x	x	x	x		x	x	x	x		
Spain		x	x	x	x	x	x	x	x	x		
Brunei Daru.	x	x		x		x		x	x	x		x
Hong Kong	x	x	x	x	x	x			x	x		
Slovenia	x	x	x	x		x		x		x		x
Sweden	x	x	x	x	x	x			x		x	

Switzerland	x	x	x	x	x	x	x	x				
Italy	x	x	x	x		x		x				
Serbia	x	x		x		x		x				x
Greece	x	x	x	x		x						
Kazakhstan	x	x	x									
Qatar	x	x		x								
United Ar. E.	x	x		x								
Total # of invariant countries	30	31	24	29	23	27	19	25	22	22	13	16

Note. Invariant Groups are marked with "x"

According to the results, items 6 (rumors) and 3 (threatened) showed the highest number of groups with invariant factor loadings while items 4 (took things) and 5 (hit/pushed) showed the lowest. These findings suggest that items 4 and 5 are likely to have a different meaning across countries and students and thus, might not be suitable indicators of bullying. Moreover, it is likely that these items are measuring other constructs not considered in the model. The number of invariant factor loadings is consistently higher among immigrant students than native students across items (except item 2) and countries for instance, factor loadings of items 6 (rumors) and 3 (threatened) were invariant for 31 and 29 countries, respectively out of 31 countries among immigrant students suggesting that item loadings are roughly equal and therefore, the bullying construct is likely to be interpreted in the same way by this population of students.

These results provide evidence supporting metric invariance for items 6 and 3. In this sense, these items could ensure valid comparisons of the latent mean of bullying across countries and within students. Moreover, results showed that the factor loadings were invariant across all the items and subpopulations of students in Belgium, Canada, Costa Rica, France, Ireland, New Zealand, United Kingdom, and the United States. Item thresholds were analyzed next. The groups for which the item thresholds were ***invariant*** are shown in Table 20.

Table 20

Summary Invariant Thresholds Bullying Scale

Country	Item 3		Item 5		Item 6		Item 4		Item 2		Item 1	
	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.
United States	x	x	x	x	x	x	x	x	x	x	x	x
Australia	x	x	x	x	x	x	x	x		x	x	x
Canada	x	x	x	x	x		x	x	x	x	x	x
Estonia	x	x	x	x	x	x	x	x		x	x	x
France	x	x	x	x	x	x	x	x	x	x		
Ireland	x	x	x	x		x		x	x	x	x	x
Luxembourg	x	x	x	x	x	x	x	x	x	x		
Macao	x	x	x	x	x	x	x	x	x	x		
New Zealand	x	x	x	x	x	x	x	x			x	x
Singapore	x	x	x	x	x	x	x	x			x	x
United Ar. E.	x	x	x	x	x	x	x	x	x			x
United King.	x	x	x	x	x	x	x	x		x	x	
Costa Rica	x	x	x	x		x	x	x		x		x
Germany	x	x		x	x	x	x	x	x	x		
Greece	x	x	x	x	x	x	x	x				x
Italy	x	x	x	x	x	x		x			x	x
Slovenia	x	x		x	x	x	x	x		x		x
Switzerland		x	x	x	x	x	x	x	x	x		
Austria	x	x	x	x	x	x		x		x		
Denmark	x	x	x	x	x	x			x	x		
Qatar	x	x	x	x		x	x	x				x
Belgium	x	x	x	x				x	x	x		
Brunei Daru.	x	x	x	x	x	x		x				
Croatia		x	x	x	x	x	x	x				

Hong Kong	x	x	x	x	x	x	x					
Kazakhstan		x	x	x	x	x	x	x				
Norway	x	x		x	x	x		x		x		
Serbia		x	x	x		x	x	x				x
Sweden	x	x		x	x	x		x		x		
Spain		x		x	x	x		x		x		
Netherlands	x		x		x					x		x
Total # of invariant countries	26	30	26	30	26	28	21	28	11	20	9	15

Note. Invariant Groups are marked with "x"

According to Table 20, thresholds of items 5 (hit/pushed) and 3 (threatened) showed the highest level of invariance both across countries and between native and immigrant students within each country, followed by item 6 (rumors). Items 1 (left out) and 2 (made fun) on the other hand, showed the lowest level of invariance across countries and immigration status. Regarding the immigration status, results showed that in general, item thresholds tend to be consistently more invariant within immigrant than native students across countries. For instance, items 3 (threatened) and 5 (hit/pushed) were invariant among immigrant students across 30 out of 31 countries, and item 6 (rumors) across 28 countries.

The United States was the only country with invariant thresholds across all items and students, followed by Australia, Canada, and Estonia which were also found to be invariant across all items and across students except for native students in item 2 (Australia and Estonia) and immigrant students in item 6 (Canada). A summary of the results from the alignment procedure in terms of number of groups per item with invariant factor loadings and thresholds is provided in Table 21.

Table 21

Number of Groups with Invariant Factor Loadings and Thresholds per Item- Bullying Scale

Item	Invariant Factor Loadings					Invariant Thresholds				
	Native	Immi.	Total	%	R^2	Native	Immi.	Total	%	R^2
6. Rumors	30	31	61	98.4	0.92	26	28	54	87.1	0.4*
3. Threatened	24	29	53	85.4	0.86	26	30	56	90.3	0.77
1. Left out	23	27	50	80.6	0.42	9	15	24	38.7	0.6
2. Made fun	22	22	44	70.9	0.04*****	11	20	31	50	0.4
5. Hit/pushed	19	25	44	70.9	0.84	26	30	56	90.3	0.69
4. Took things	13	16	29	46.7	0.72	21	28	49	79	0.38

*****According to a post by Asparouhov on Mplus Discussion (2018) the R^2 can be close to zero even for invariant items when the power was not sufficient to establish the non-invariance (e.g., small sample sizes, empty cells in bivariate tables) or when the average aligned loading is close to zero.

Regarding the factor loadings, results indicate that items 6 (rumors) and 3 (threatened) showed the highest level of invariance across groups. Specifically, the factor loadings from item 6 were invariant across 61 groups out of 62, and item 3 across 53 groups. Item 6 was also equally invariant across native and immigrant students. Factor loadings of item 4 (took things) on the other hand, showed the lowest level of invariance across the groups. Specifically, factor loadings were invariant across 29 out of 62 groups under analysis suggesting that this item is likely to have a different meaning across cultural groups and thus, might not lead to valid comparisons of the latent means across cultural groups.

In terms of item thresholds, results show that the thresholds of items 3 (threatened) and 5 (hit/pushed) had the highest level of invariance across the 62 groups under analysis (90.3%) followed by item 6 with 87.1% of invariant groups. On the other hand, items 1 (left out) and 2 (made fun) showed the lowest level of invariance across the groups (38.7 and 50%, respectively).

The alignment procedure also provides an R^2 measure that indicates the degree of invariance for the parameter under analysis. According to Muthén and Asparouhov (2018), it shows how much of the variation in the configural parameter across the groups can be explained by variation in the factor means and variances so that high values suggest a high degree of measurement invariance. According to the results from table 21, the factor loadings of items 6 (rumors), 3 (threatened), and 5 (hit/pushed) are highly invariant across the groups under analysis so that 92, 86 and 84% of their variation in the configural model can be explained by the variation in the mean and variance of bullying, respectively. However, the factor loadings of items 2 (made fun) and 1 (left out) were the least invariant so that their variation accounts for less than 1% and 0.42% of the variation in bullying, respectively.

Regarding the thresholds, results show that the thresholds of item 3 (threatened) are the most invariant across countries and that 80% of its variation across the groups in the configural model can be explained by the variation in the mean and variance of bullying across groups. Thresholds of items 3 (threatened) and 5 (hit/pushed) also showed a high level of invariance so that 77% and 69% of their variation can be explained by the variation in the factor mean and variance.

The thresholds of item 4 (took things) on the other hand, showed the highest level of non-invariance in that only 38% of the variation in the item thresholds can be explained by the variation in bullying. Overall, these findings suggest items 6 (rumors) and 3 (threatened) are more likely to lead to reliable comparisons of bullying across countries and students with different immigration status. Moreover, the results provided an average invariance index which is a general score of metric and scalar invariance and

can take values from 0 to 1 where 1 indicates perfect scalar invariance. The index for the bullying scale was 0.59 suggesting that the means can be meaningfully compared across the groups with 59% of confidence.

However, Asparouhov and Muthén (2014) suggested a rule of thumb where a limit of 25% non-invariance is safe for trustworthy alignment results. Even though the average percentage of noninvariant factor loadings was 24.4%, the average percentage of noninvariant thresholds for the items in the bullying scale were higher than 25% (27.4 percent) therefore, there is not enough evidence of trustworthy alignment results for the bullying scale across countries and students (Muthén & Asparouhov, 2014).

4.2.2.2 Sense of Belonging at School. Measurement invariance of the six items from the sense of belonging at school scale was also assessed through the alignment optimization procedure. As with the bullying scale, a total of 62 groups identifying countries and immigration status were included in the analysis. The model for the sense of belonging at school scale was shown in Figure 8 and the Mplus code for this analysis was the same as that used for the bullying scale except for the group that was included as baseline which was that of immigrant students living in United Kingdom (labeled as 82600). Table 22 shows the groups with ***invariant*** factor loadings.

Table 22

Summary Invariant Factor Loadings Sense of Belonging at School Scale

Country	Item 4		Item 2		Item 6		Item 1		Item 5		Item 3	
	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.
Belgium	x	x	x	x	x	x	x	x	x	x	x	x
Brunei Daru.	x	x	x	x	x	x	x	x	x	x	x	x
Canada	x	x	x	x	x	x	x	x	x	x	x	x
Denmark	x	x	x	x	x	x	x	x	x	x	x	x
Ireland	x	x	x	x	x	x	x	x	x	x	x	x
Italy	x	x	x	x	x	x	x	x	x	x	x	x
Luxembourg	x	x	x	x	x	x	x	x	x	x	x	x
Macao	x	x	x	x	x	x	x	x	x	x	x	x
New Zealand	x	x	x	x	x	x	x	x	x	x	x	x
Singapore	x	x	x	x	x	x	x	x	x	x	x	x
United King.	x	x	x	x	x	x	x	x	x	x	x	x
United States	x	x	x	x	x	x	x	x	x	x	x	x
Australia	x	x		x	x	x	x	x	x	x	x	x
Costa Rica	x	x	x	x	x	x	x	x	x	x		x
Estonia	x	x	x	x	x	x	x	x		x	x	x
Germany	x	x	x	x	x	x	x	x		x	x	x
Hong Kong	x	x	x	x	x	x	x	x	x	x		x
Switzerland	x	x	x	x	x	x	x	x	x	x		x
Croatia		x		x	x	x	x	x	x	x	x	x
France	x	x	x	x	x	x			x	x	x	x
Greece	x	x	x	x	x	x		x	x	x		x
Slovenia		x	x	x	x	x	x	x	x	x		x
Netherlands	x	x	x	x		x		x		x	x	x
Serbia		x		x	x	x		x	x	x	x	x

United Ar. E.	x	x		x	x	x	x	x		x		x
Norway	x		x	x	x			x	x	x		x
Spain	x	x	x	x		x		x		x		x
Austria	x	x	x	x		x		x				x
Kazakhstan				x		x	x	x	x	x		x
Sweden	x	x	x	x		x		x				x
Qatar				x			x			x		x
Total # of invariant countries	26	28	25	31	25	29	23	29	23	29	19	31

Note. Invariant Groups are marked with "x"

The factor loadings of the six items from the sense of belonging at school scale seem to be highly invariant across countries and students except for item 3 (feel like I belong) that showed the lowest number of invariant loadings among native students. The number of invariant loadings is consistently higher among immigrant students across all the items. Moreover, the countries where all the items were invariant across immigrant and native students include Belgium, Brunei Darussalam, Canada, Denmark, Ireland, Italy, Luxembourg, Macao, New Zealand, Singapore, United Kingdom, and the United States.

Overall, this finding provides initial evidence about the invariance of the scale suggesting that the items are suitable indicators to measure the target construct and that their content is likely to be interpreted in a similar way across cultural groups. Therefore, comparisons of the factor mean across the groups are expected to lead to trustworthy inferences based on test scores. The level of invariance of the thresholds was evaluated next and results are shown in Table 23.

Table 23

Summary Invariant Thresholds Sense of Belonging at School Scale

Country	Item 4		Item 6		Item 1		Item 2		Item 3		Item 5	
	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.	Native	Imm.
Australia	x	x	x	x	x	x	x	x	x	x	x	x
Canada	x	x	x	x	x	x	x	x	x	x	x	x
Qatar	x	x	x	x	x	x	x	x	x	x	x	x
New Zealand	x	x	x	x	x	x	x	x		x	x	x
Singapore	x	x	x	x	x	x	x	x	x	x		x
United Ar. E.	x	x	x	x	x	x	x	x	x	x		x
Serbia	x	x		x	x	x	x	x		x	x	x
United States	x	x		x		x	x	x	x	x		x
Greece	x	x	x	x	x	x		x		x		x
Ireland	x	x		x	x	x		x		x	x	x
Slovenia		x	x	x	x	x	x	x	x	x		
Switzerland	x	x	x	x	x	x		x			x	x
Brunei Daru.		x		x	x	x		x	x	x		x
Netherlands	x	x	x	x		x		x		x		x
Norway	x	x		x			x	x	x	x		x
United King.	x	x		x	x	x		x			x	x
Germany	x	x		x	x	x				x		x
Hong Kong				x	x	x	x	x	x	x		
Macao	x	x	x	x	x		x	x				
Costa Rica	x	x		x		x		x				x
Croatia		x		x			x	x	x	x		
France	x	x		x			x	x				x
Kazakhstan		x		x			x	x	x	x		
Sweden		x	x	x		x		x		x		

Belgium	x	x		x		x		x				
Estonia		x	x	x					x	x		
Italy		x	x	x		x		x				
Spain	x	x	x	x						x		
Austria		x	x	x						x		
Denmark		x	x	x				x				
Luxembourg		x	x	x				x				
Total # of invariant countries	20	30	18	31	16	21	15	27	13	22	8	18

Note. Invariant Groups are marked with "x"

According to Table 23, the thresholds from item 4 (feel awkward/out of place) showed the highest level of invariance across countries and students. Specifically, thresholds of item 4 were invariant across immigrant students living in 30 out of 31 countries, and across native students living in 20 countries. Whereas item 5 (students like me) showed the lowest level of invariance across countries and students so that only 26 thresholds were invariant across the 62 groups.

Results also showed that the number of invariant thresholds was consistently higher among immigrant students than their native peers across all the items. Moreover, Australia, Canada, and Qatar were the countries with the highest number (12 out of 12) of invariant thresholds across items while Austria, Denmark, and Luxembourg had the lowest (4 out of 12).

The findings regarding item loadings and thresholds provide evidence of metric invariance since the item loadings appear to be the same across the groups, however, there does not seem to be evidence of scalar invariance since invariance does not hold for several item thresholds. In this sense, results suggest that even though a one-unit increase in the construct is likely to mean the same across countries and groups of students per immigration status, students with the same level of sense of belonging at school will not have the same expected response on the scale because their responses will probably depend on their immigration status and the country where they reside. Given this finding, there is not sufficient evidence to ensure trustworthy cross-cultural comparisons of the latent mean. The summary of the results from the alignment procedure is shown in Table 24.

Table 24

Number of Groups with Invariant Factor Loadings and Thresholds per Item- Belong Scale

	Invariant Factor Loadings					Invariant Thresholds				
Item	Native	Immi.	Total	%	R^2	Native	Immi.	Total	%	R^2
2. Make friends easily	25	31	56	90.3	0.2*	15	27	42	67.7	0.2
4. Awkward/out of place	26	28	54	87.1	0.1*	20	30	50	80.6	0.96
6. Feel lonely	25	29	54	87.1	0.0*	18	31	49	79	0.93
1. Feel/outsider	23	29	52	83.9	0.0*	16	21	37	59.7	0.8
5. Students like me	23	29	52	83.9	0.28*	8	18	26	42	0.54
3. Feel belong	19	31	50	80.6	0.33*	13	22	35	56.5	0.2

*According to a post by Asparouhov on Mplus Discussion (2018) the R^2 can be close to zero even for invariant items when the power was not sufficient to establish the non-invariance (e.g., small sample sizes, empty cells in bivariate tables) or when the average aligned loading is close to 0.

Table 24 shows that even though most factor loadings were found invariant, the R^2 values are very low suggesting that the variation in the factor loadings cannot be fully explained by the variation in the latent construct and thus, it is probably due to cultural-related differences in the groups. In this sense, even though the items could be pointing to measure sense of belonging at school, the contribution of each item in terms of loadings' weights could be different across the groups.

Regarding the item thresholds, results indicate that items 4 (awkward/out of place) and 6 (feel lonely) showed the highest percentage (80.6% and 79%, respectively) of invariant thresholds across the groups suggesting that the psychometric features of these items are likely to hold across groups therefore, these items could be used to make trustworthy comparisons across countries and students. Thresholds from items 5 (students like me) and 3 (feel belong) on the other hand, showed the lowest percentage of invariance across the groups suggesting that answers to these items might not reflect the level of the respondents on the target construct but other cultural-related variables of the respondents instead moreover, the answers could be reflecting translation-related issues. Thus, these items might not be suitable to be used to comparison purposes.

In terms of the R^2, results showed that items 6 (feel lonely) and 4 (awkward/ out of place) account for 93% and 96% of the variation in the mean and variance of the target construct whereas only 20% of the variation in the thresholds of items 2 and 3 can be explained by variation in sense of belonging at school. Thus, these items showed the highest level of non-invariance and are likely to reflect differences in cultural-related variables rather than differences in the latent construct.

Overall, the findings suggest items 6 (feel lonely) and 4 (awkward/out of place) showed the highest level of invariance and are suitable indicators of sense of belonging at school across countries and students with different immigration status. Regarding the average invariance index, results showed that the index for the sense of belonging at school scale was 0.39, a value close to 0 suggesting that the means of the latent construct can be compared across the groups with only 39% of confidence.

On the other hand, according to the rule of thumb by Asparouhov and Muthén (2014) results indicate that the average percentage of noninvariant factor loadings and thresholds for the items in the sense of belonging at school scale are close to 25% (14.5% and 35.7%, respectively for an average of 25.1%) thus, there is not enough evidence of trustworthy alignment results for the sense of belonging at school scale across countries and students.

4.2.2.3 Reading Literacy. Measurement invariance was assessed at the test level where the average scores across the ten plausible values reported for each of the three subscales were used as indicators of reading literacy. As with the non-cognitive measures, 62 groups were also analyzed to account for country and immigration status. The model that was analyzed included three indicators which in this case correspond to the three subscales and one latent factor (reading literacy) as was shown in Figure 9.

The analyses were performed in Mplus following the same code that was used for the non-cognitive scales. Thus, the analyses began with the configural model with means and variances fixed to one and zero, respectively and factor loadings and intercepts were freely estimated. Free alignment was implemented first to identify the baseline group that would prevent model misspecification and which according to the results should be

the group of immigrant students living in Denmark (labeled as 20800) therefore, the fixed alignment was then implemented following this suggestion.

The groups with ***invariant*** factor loadings according to the alignment optimization procedure are shown in Table 25.

Table 25

Summary Invariant Factor Loadings Reading Literacy Scale

Country	Understand		Evaluate and Reflect		Locate Information	
	Native	Imm.	Native	Imm.	Native	Imm.
Austria	x	x	x	x	x	x
Sweden	x	x	x	x	x	x
Switzerland	x	x	x	x	x	x
United Ar. E.	x	x	x	x	x	x
Belgium		x	x	x	x	x
Costa Rica		x	x	x	x	x
Greece	x	x	x	x	x	
Hong Kong		x	x	x	x	x
Netherlands	x	x		x	x	x
New Zealand	x	x			x	x
United Kingdom	x	x		x	x	x
Estonia		x	x	x		x
Ireland	x	x		x		x
Italy		x	x	x		x
Luxembourg	x	x		x		x
Spain	x	x	x	x		
United States	x	x		x		x
France	x	x				x
Kazakhstan		x		x		x
Macao		x	x		x	
Norway		x		x		x
Brunei Daru.	x		x			
Denmark					x	x
Germany	x	x				
Qatar		x		x		
Serbia		x				x
Singapore	x	x				
Australia		x				
Canada			x			
Croatia			x			
Slovenia	x					
Total # of invariant countries	17	26	15	19	13	20

Note. Invariant Groups are marked with "x"

According to the results, the subscale *understand* showed the largest number of invariant factor loadings across countries and students. Specifically, the factor loadings were invariant among native students across 17 countries out of 31 and among immigrant

students across 26 countries. The indicator of *locate information* on the other hand showed the lowest level of invariance so that factor loadings were invariant among native students across 13 countries and among immigrant students across 20 countries moreover, the number of invariant factor loadings was consistently higher across the three indicators among immigrant students than their native peers. This finding could be suggesting that immigrant students from different countries seemed to respond similarly to the items.

Results also show that the factor loadings were invariant across all the indicators and students in Sweden, Switzerland, and the United Arab Emirates whereas Australia showed only one invariant factor loading. Overall, these findings suggest that the indicators of reading literacy are likely to be interpreted in a different way across countries and students, particularly non-immigrant students. Intercepts were evaluated next. Results are shown in Table 26.

Table 26

Summary Invariant Intercepts Reading Literacy Scale

Country	Locate Information		Understand		Evaluate and Reflect	
	Native	Imm.	Native	Imm.	Native	Imm.
Australia	x	x	x	x	x	
Canada	x	x		x	x	x
Hong Kong	x	x	x		x	x
Qatar	x		x	x	x	x
United States	x	x	x	x	x	
Belgium		x	x		x	x
Brunei Daru.		x	x	x		x
Denmark	x	x		x		x
Germany		x	x	x		x
Spain	x	x			x	x
United King.		x		x	x	x
Greece	x	x				x
Italy		x			x	x
Macao		x	x	x		
Norway			x	x		x
Sweden			x	x	x	
Austria	x	x				
Kazakhstan	x	x				
Netherlands			x	x		
New Zealand	x			x		
Serbia	x	x				
Singapore		x				x
Costa Rica		x				
Croatia		x				
Estonia				x		
France				x		
Luxembourg	x					
Slovenia		x				
United Ar. E.			x			
Ireland						
Switzerland						
Total # of invariant countries	13	20	12	15	10	13

According to Table 26, the indicator of locate information showed the largest number of invariant intercepts across countries and students whereas the indicator of evaluate and reflect showed the lowest (33 and 23 out of 62, respectively). Regarding the

immigration status, results show that the number of invariant intercepts was consistently higher among immigrant than native students across all the indicators and none of the countries had invariant thresholds across all items and countries. In general, results show a low number of invariant intercepts for the reading literacy subscale suggesting that the meaning of the indicators does not hold across countries and students. The summary of the alignment procedure in terms of number of groups per indicator with invariant factor loadings and intercepts is shown in Table 27.

Table 27

Number of Groups with Invariant Factor Loadings and Intercepts per Indicator- Reading Scale

	Invariant Factor Loadings					**Invariant Intercepts**				
Indicator	Native	Immi.	Total	%	R^2	Native	Immi.	Total	%	R^2
2. Understand	17	26	43	69.4	0.0*	12	15	27	43.5	0.13
3. Evaluate/ reflect	15	19	34	54.8	0.0*	10	13	23	37.1	0.24
1. Locate Info.	13	20	33	53.2	0.0*	13	20	33	53.2	0.22

*****According to a post by Asparouhov on Mplus Discussion (2018) the R^2 can be close to zero even for invariant items when the power was not sufficient to establish the non-invariance (e.g., small sample sizes, empty cells in bivariate tables) or when the average aligned loading is close to zero.

In terms of factor loadings, results show that the indicator of understand had the highest number of invariant factor loadings (69%) specifically, the factor loadings were invariant cross 43 groups out of 62, while the indicator for locate information showed the lowest (53%). As previously mentioned, the factor loadings were consistently more invariant among immigrant than native students. Regarding the intercepts, results show that in general, the number of invariant intercepts was low, intercepts were invariant in less than half of the total number of countries under analysis. Intercepts of the indicator

for locate information showed the highest level of invariance across 33 countries which is still low.

Finally, in terms of the R^2 measure that indicates the degree of invariance for the parameters under analysis, the values were very low (below 0.25) for both factor loadings and intercepts suggesting that the variation in the parameters are not fully accounting for the variation in the latent construct thus, it is likely that the variation in the parameters is due to other cultural-related constructs not considered in the model.

The overall findings do not provide enough evidence of measurement invariance (metric or scalar) for the reading literacy scale, which is also confirmed by the average invariance index, which was equal to 0.09, indicating that the latent means cannot be meaningfully compared across the groups.

4.2.3 Evaluation of the Relationship between the Non-cognitive Measures and the Performance on Reading literacy

The latent structural regression model shown in Figure 8 was analyzed next. The following aspects were considered for identification purposes:

1. Unit loading identification constraints were used to scale the factors in a metric similar to the metric of the common variance of the reference variable. The first factor loading for every common factor was fixed to 1.0.
2. The common factors were standardized by fixing the variances to 1.0.
3. Each factor had at least three indicators.

Results for the latent structural equation model are shown in Table 28.

Table 28

Summary Fit Statistics Latent Structural Equation Model

Model Fit Statistics	Estimates
Chi-Square X^2	124437.9 (*df*=87, *p*=0.000)
RMSEA	0.081, 90% *CI* [0.081, 0.081]
CFI	0.95
TLI	0.94
SRMR	0.054

The Chi-square test was used to evaluate the null hypothesis stating that there are no statistically significant differences between the covariances predicted by the model and the population covariance matrix. Given that the results indicated statistical significance (*p*= 0.000), the null hypothesis was rejected suggesting that the model under analysis might not be an accurate representation of the underlying relationships among the latent constructs. The chi-square statistic is sensitive to sample size and trivial differences are likely to be flagged when the sample size is large specifically, large samples lead to smaller model-data discrepancies that in turn, lead to rejection of the exact-fit (null) hypothesis as it is the case in this analysis thus, the evaluation of model fit will not solely rely on this statistic (Kline, 2016).

Regarding the absolute fit index, Root Mean Square Error of Approximation (RMSEA) results showed a value close to zero which indicates acceptable fit since this index is a badness-of-fit statistic where a value of zero corresponds to the best result. Following the guidelines by Isac et al. (2019), values of RMSEA between 0.08 and 0.010 are indicators of acceptable fit. In this sense, this finding provides evidence in favor of the hypothesized model suggesting that the relationships stated in the model might be an accurate representation of the true relationships among the latent constructs.

The incremental comparative fit index (CFI) was also obtained. This is a goodness-of-fit statistic, its values range from 0 to 1.0 where 1.0 indicates perfect fit moreover, Byrne and Van der Vijver (2010) and Isac et al. (2019) suggest that a value of 0.95 serves a rule of thumb cut point of acceptable fit. Results show a value higher than 0.9 providing evidence in favor of the hypothesized model. Specifically, the fit of the hypothesized model is 95% better than that of the baseline model. Similarly, the Tucker-Lewis index (TLI) also showed a value higher than 0.9. This index is also a goodness-of-fit statistic that imposes greater penalty for model complexity and based on the results, it offers evidence favoring the hypothesized relationships among the latent variables under analysis.

Finally, the standardized root mean square residual (SRMR), a measure of the average standardized covariance residual was also obtained. This statistic is a badness-of-fit indicator where a value of zero corresponds to perfect fit. Results show a value of 0.05 suggesting similarity between the observed and predicted residual correlations thus, this is another evidence in favor of a well-fitting model. Estimates for the model are presented in Table 29.

Table 29

Summary Standardized Results Latent Structural Equation Model

Latent Factors	**Estimate**	**SE**	**Two-tailed p-value**
Exposure to bullying	-0.23	0.002	0.000
Sense of Belonging at School	0.011	0.003	0.000
Correlation bullying/ sense of belonging	-0.4	0.002	0.000

The two latent factors -exposure to bullying and sense of belonging at school- are significant ($p < 0.001$) predictors of reading literacy. Moreover, the relationship between

exposure to bullying and reading literacy was negative suggesting that high values in the bullying construct can lead to low performance in reading literacy whereas the relationship between sense of belonging at school and reading literacy was positive indicating that high values in sense of belonging at school lead to high performance in reading literacy. Regarding the correlation between the latent constructs, results show a negative and significant ($p < 0.001$) correlation so that high values on bullying are related to low values on sense of belonging at school. Even though the correlation was significant, the estimate was rather low (-0.4) indicating a moderate correlation.

Finally, the residuals, that is, the differences between the observed and predicted covariances, were also inspected. For this analysis, the correlation residuals were inspected where values equal or close to zero are indicators of good fit whereas values higher than 1.0 suggest poor fit. Specifically, positive correlations higher than 1.0 indicate that a common variable not considered in the model might be affecting the two indicators in the same direction whereas negative values indicate that as one variable increases the other decreases due to the unconsidered variable. Results are shown in Table 30.

Table 30

Correlation Residuals

	Sense of belonging at school						Bullying						Reading		
	BE1	BE2	BE3	BE4	BE5	BE6	BU1	BU2	BU3	BU4	BU5	BU6	LOC	UN	EV
BE1	0														
BE2	-0.057	0													
BE3	-0.054	**0.188**	0												
BE4	0.034	-0.076	-0.027	0											
BE5	-0.061	**0.242**	**0.176**	-0.085	0										
BE6	0.036	-0.044	-0.07	0.045	-0.055	0									
BU1	**-0.145**	-0.082	-0.083	-0.091	**-0.111**	**-0.128**	0								
BU2	-0.077	-0.023	-0.042	-0.039	-0.064	-0.049	**0.125**	0							
BU3	0.009	0.052	0.019	0.037	-0.014	0.033	-0.022	-0.028	0						
BU4	0.028	0.057	0.036	0.057	-0.001	0.049	-0.058	-0.039	0.017	0					
BU5	0.03	0.065	0.035	0.055	0.006	0.052	-0.068	-0.03	0.024	0.063	0				
BU6	-0.031	0.024	-0.018	0.002	-0.047	-0.014	0.048	0.037	-0.009	-0.022	-0.025	0			
LOC	0.024	-0.087	-0.003	0.027	0.019	-0.006	0.038	**0.105**	-0.04	-0.026	-0.024	-0.005	0		
UN	0.024	-0.09	0.002	0.026	0.018	-0.011	0.039	**0.103**	-0.04	-0.022	-0.023	-0.007	0	0	
EV	0.018	-0.086	0.003	0.021	0.023	-0.014	0.043	**0.11**	-0.031	-0.014	-0.014	-0.002	0	0	0

Most residuals are close to zero except for ten residuals (bolded and highlighted) that showed values higher than 0.1 which could be viewed as possible evidence of poor local fit. Specifically, results show high and *positive* correlations among three indicators from the sense of belonging at school scale (items 2, 3, and 5) suggesting that the hypothesized model could be underpredicting the observed associations among these indicators that is, the model could be underpredicting the relationship between (a) indicators two and three by 0.18, (b) indicators two and five by 0.24, and (c) indicators three and five by 0.17. Given the statements presented in these indicators ("I make friends easily at school", "I feel like I belong at school", "other students seem to like me"), it is likely that they are highly correlated. Likewise, the residuals of the indicators one ("other students left me out of things on purpose") and two ("other students made fun of me") from the bullying scale were positively correlated indicating that the hypothesized model could also be underpredicting the relationship between these indicators by 0.12.

The residuals of the remaining indicators of the sense of belonging at school scale ("I feel like an outsider at school," "I feel awkward and out of place in my school," "I feel lonely at school") also showed high *negative* correlations with the residuals from the first indicator of the bullying scale ("other students left me out of things on purpose"). This finding suggests these indicators could be measuring the two latent factors that is, sense of belonging at school and bullying, and that the indicators might be sharing some variance (shared error variance) that is unique to them but irrelevant to the target latent constructs. Finally, the residuals of the second indicator from the bullying scale showed

positive correlations with the residuals of all the indicators from the reading scale thus, it is likely that these indicators might be sharing error variance and the model is underpredicting these construct-irrelevant relationships.

Taken together, the overall findings support the hypothesized relationships among the latent variables as measure by PISA 2018. Specifically, exposure to bullying and sense of belonging at school are related to one another and these two constructs impact the performance on the reading literacy scale.

CHAPTER

V DISCUSSION

International large-scale educational assessments (ILSAs) have played a relevant role in educational policies across countries. In fact, results from these assessments can be used as a standard against which countries evaluate the performance of their educational systems. In this context, ILSAs can have a significant impact in educational systems throughout the world as their results are used by governments as input for decision-making purposes.

Given the potential impact that ILSAs can have, the psychometric features of these assessments must be carefully assessed and empirical evidence about the extent to which the inferences made based on test results are valid must be collected. To do so, the first step is to determine if the test results have the same meaning across countries and groups of examinees that is, if the measures are invariant so that results can be compared directly among countries.

However, the evaluation of measurement invariance is usually problematic when the groups under analysis differ from one another in terms of cultural and socio-economic characteristics as it is the case of ILSAs that are to be compared across countries. Moreover, the task becomes even more challenging when the student population within each country is diverse and includes not only native but immigrant students. Despite the challenges, ILSAs provide valuable information and in order to make the best use of the results, evidence-based modeling techniques that can handle the features of these assessments and populations must be implemented.

The general purpose of this dissertation was to provide evidence about the extent to which the 2018 Programme for International Student Assessment (PISA) provides invariant measures of reading literacy, exposure to bullying, and sense of belonging at school for immigrant students from diverse cultural and linguistic backgrounds across the countries that host large populations of immigrants. Moreover, given that test performance can be impacted by non-cognitive variables, the constructs exposure to bullying and sense of belonging at school were analyzed as potential predictors of student performance in reading literacy.

Two statistical techniques were implemented to evaluate the extent to which the measures from PISA 2018 were invariant across countries and students with different immigration status (native vs. immigrant). Three scales were selected for this analysis: reading literacy, which was the main subject of PISA 2018, and two non-cognitive scales sense of belonging at school and exposure to bullying.

The overall results showed that the alignment optimization procedure was a more suitable statistical tool than the traditional modeling technique -multiple-group confirmatory factor analysis- for the evaluation of measurement invariance when the data under analysis are collected through ILSAs since it can handle the features and complexities of these data while allowing for the incorporation of the immigration status into the analysis. Moreover, the alignment optimization provides more informative results that can guide test users as to the best ways to use the results.

Regarding the invariance of the measures, results provided evidence supporting the invariance of most of the items from the exposure to bullying scale, partial invariance

was also achieved for the sense of belonging at school scale however, the results did not provide evidence of the measurement invariance for the reading literacy scale.

The most salient findings from this dissertation were:

- The lack of invariance in the sense of belonging at school scale is probably related to (a) the fact that some items are negatively worded, (b) the sensitivity of the Likert-scale to cultural differences, and (c) the cultural differences among countries where some tend to reinforce this construct more than others.
- The number of invariant item parameters was larger among immigrant students than among their native peers suggesting that immigrant students across the world might be experiencing similar situations that leads them to interpret the test items in a similar way.
- There was no evidence of measurement invariance for the reading literacy scale probably due to the high sensitivity of this construct to cultural and language differences.

In the next sections of this discussion chapter, I delve further into the specifics of these findings, focusing on those of most significance and practical consequences.

5.1 Multiple-Group Factor Analysis (MGCFA)

5.1.1 Exposure to Bullying Scale

The analysis began with the evaluation of the configural model where the fit statistics provided evidence in favor of the configural model suggesting that the bullying construct is being measured in the same way across countries thus, it is likely that examinees across countries used the same conceptual framework when they answered the

test items. However, results showed that Italy, Spain, and Canada were the countries that contributed the most to the chi-square suggesting that these countries could be deteriorating model fit in terms of measurement non-invariance for the configural model. This finding provides initial evidence suggesting that there might be internal educational dynamics that could be leading students with the same exposure to bullying to show a different performance in the bullying scale which in turn could be reflecting disparities in the educational experiences among these countries.

On the other hand, and despite the acceptable fit of the configural model, the inspection of correlations between residuals indicated problems with the scale. Specifically, the inspection indicated that items 1 and 5 ("left out" and "took/destroyed things", respectively) were probably related to one another in 13 out of the 31 countries under analysis suggesting that there might be some circumstances in those countries that could be impacting how the situations expressed in these two items are experienced. The students' experience of being left out by their peers could be due to differences in the inclusion practices or cultural appropriation across educational systems. As mentioned by Samara et al. (2019), the interpretation of the types of bullying varies across countries and the instruments used to measure bullying are usually sensitive to socio-cultural differences such as cultural norms, socioeconomic inequality, and cultural values thus, these findings could be reflecting social inequalities.

The number of countries that were flagged based on the inspection of correlations between residuals increased as more equality restrictions were imposed in the model parameters so that 24 and 29 countries out of 31 were flagged in the metric and scalar models, respectively. This finding indicates that measurement invariance decreased as

equality constraints increased. Therefore, it is likely the unit of measurement of the latent variable (exposure to bullying) varies across countries meaning that observed variation in the construct might be reflecting variation due to group membership and if so, direct comparisons of test scores across countries could reflect the impact of variables other than the latent construct.

Regarding the metric and scalar models, no sound evidence was found to support measurement invariance. In fact, model fit was increasingly deteriorated as equality constraints increased and the change in the incremental fit statistics suggested that constrained models were statistically worse than the less restrictive models. These findings indicated not only that the latent construct is probably expressed differently across countries but also that students from different countries might be using the response scale in a different way. There are two elements worth noting to this regard. First, the wording of the instructions about the response scale could lead to confusion because the response options are targeting different levels of frequency (never, almost never, etc) but in the instructions students are asked to provide answers based on the experiences they have had 12 months prior to the test. Students who did not experience any of the situation stated in the test items 12 months prior to the test but might be experiencing situations recently might choose not to report any situation. Instructions provided to test takers are as important or perhaps even more important than the test items especially when the target population is as diverse as the student population participating in ILSAs. Comprehension issues could easily arise when examinees are taking the test in a foreign language and even when they are undergoing stressful situations (e.g., the experience of being an immigrant and all that it entails) thus, it is

important that test developers and researchers combine efforts to find the simplest way to deliver test instructions in order to control for potential biases related to comprehension.

The findings from the MGCFA analysis could also be reflecting psychological differences among students specifically between immigrant and native students within each country. As pointed by Duong et al. (2016) and Rosen et al. (2013), immigrant students are more likely to be exposed to peer aggression than their native peers and thus, are in higher risk of undergoing health and adjustment problems including severe anxiety, stress, depression, engagement in risk behaviors such as aggression, delinquency, and substance abuse among others, which in turn has an ultimately negative impact on their overall test performance. Therefore, the differences in terms of inclusion policies targeting immigrant students can result in the lack of measurement invariance across countries so that some students are more likely to experience general psychological distress related to their status as immigrants and this individual state can have a direct impact in the way they approach the test items moreover, depending on the severity of the psychological distress, it is also possible that some test items could trigger stress responses. In this sense, test developers are also urged to consider these scenarios from the initial stages of test development through the implementation of evidence-based qualitative methodologies that allow them to better understand the most common experiences of immigrant students so that they can control for biases related to the psychological state of the students.

Another finding that is worth noting was that three countries were systematically flagged in the metric and scalar models for bullying due to their high contribution to the deterioration of the Chi-square statistic: Spain (which was also flagged in the configural

model), Kazakhstan, and Brunei Darussalam. According to the cultural model by Hofstede (2011) both Spain and Kazakhstan are similar to one another in terms of the cultural dimension denoted as *uncertainty avoidance* which refers to the extent to which the citizens of a country feel threatened by uncertainty and have created institutions or rules to avoid it. Spain and Kazakhstan have a high score (86 and 88, respectively) in this dimension suggesting that people are likely to have a negative attitude towards change which can also promote intolerance. In this scenario, it is very likely that the educational systems promote the same beliefs through their daily dynamics and thus, intolerant behaviors could lead to bullying-related behaviors against students who do not strictly observe the rules or other expected behaviors. Even though, Brunei Darussalam is not considered in the Hofstede's tool to compare countries, it could be inferred that this country could have a similar score in this dimension since its government in an absolute constitutional monarchy.

This finding overall could be suggesting that countries with a high tendency towards uncertainty avoidance could have particular internal educational dynamics that can impact the overall experience of being exposed to bullying thus, further inspection of this characteristic is encouraged to identify potential confounding variables and control their impact in educational and psychological international measures.

In summary and regarding the first purpose of this dissertation, the results from the analysis on the bullying scale indicate that even though the items seem to be targeting the latent construct "exposure to bullying", there is no sound evidence from MGCFA about the invariance of the exposure to bullying scale across countries. No evidence was

collected about the extent to which this scale is invariant across immigrant and native students due to the statistical features of the MGCFA modeling technique.

However, two comments should be made in the light of these findings. First, one of the major methodological challenges when developing international assessments is related to translation equivalence. According to Samara et al. (2019), the cultural systems determine the meaning and characteristics of psychological constructs and cognitive processes leading to methodological biases such as translation biases, perception of response styles, lack of familiarity with testing procedures, and construct underrepresentation and these in turn, can have a direct impact on test items or even on the whole measurement instrument violating the requirements of measurement invariance while hindering the generalizability of the target construct.

And second, the response style is also likely to be influenced by social desirability and this can be particularly likely in the case of measure of aggression given that in most cultures aggressive behavior is highly undesirable. As noted in a study by Vigil-Colet et al. (2012), self-reported aggression is considerably reduced among examinees with high levels of social desirability leading to biased test responses. Therefore, it is recommended that test developers and testing companies in general, account for social desirability-related biases when developing psychological measures targeting sensitive constructs such as bullying. Test developers could implement alternate item formats (e.g., presenting hypothetical situations) to prevent biased responses that can deteriorate the validity of the interpretations to be made based on test scores.

5.1.2 Sense of Belonging at School Scale

The analyses on the sense of belonging at school scale did not provide evidence of invariance however, they did provide useful information that helped contextualize and more importantly explain to some extent the lack of invariance.

The evaluation of measurement invariance began with the configural model where the results did not provide strong evidence in favor of the less restrictive model moreover, all the countries displayed problems with the correlations between residuals suggesting that the model was both over and underpredicting the sample polychoric correlations.

The evaluation of the configural model also showed that the countries that made the largest contribution to the Chi-Square were Qatar, Kazakhstan, and United Arab Emirates. According to the cultural dimensions by Hofstede (2011), these countries have almost the same level of individualism (scores of 25, 20, and 25, respectively) which indicates that they are highly collectivistic societies where strong relationships based on interdependence are encouraged and reinforced. In this scenario the sense of belonging can be easily promoted and shared among the members of the society which in turn influences their reported level of sense of belonging at school.

Qatar and United Arab Emirates were also the countries with the highest number of flagged residual correlations. This finding confirms that the configural model does not properly reflect the relationships between the test items and the latent construct meaning that students within these countries are interpreting the latent construct in a different way than the other countries.

The results also indicated that the residual of item 2 (“I make friends easily”) was repeatedly flagged across countries therefore, this item can be particularly problematic to the establishment of measurement invariance. As previously mentioned, psychological-related constructs are more sensitive to cultural differences among test takers and this is likely to be the problem with this item moreover, these items are also likely to be subjected to social desirability especially among adolescents which can also become a source of item bias.

On the other hand, and regarding the metric model (that provides information about the way in which test takers are using the response categories), results showed that Spain, United Arab Emirates, and Kazakhstan made the largest contribution to the Chi-Square. As previously mentioned, these countries have cultural similarities that can impact the educational dynamics and the way students experience the sense of belonging at school. Furthermore, according to the results there was no evidence in support of the metric model: severe misspecification was found in the evaluation of residuals and the overall fit did not lead to improvement over the configural model which is why the evaluation did not proceed to the scalar model.

The findings from the configural model suggest that residual item correlations could be pointing cultural differences across countries in terms of individualism/collectivism. This is because as pointed by Meng et al. (2018), eastern and western cultures are known for their complexities and marked differences especially in terms of collectivism and individualism since individualist cultures focus on the pursuit of personal goals, autonomy, and independence from others, whereas collectivist cultures focus on the preservation of relationships, social harmony, and independence.

More importantly, these cultural features determine the individual experiences of students as well as the way in which they respond to test items specially those intended to measure non-cognitive constructs (Meng et al., 2018). For instance, collectivistic countries that emphasize interdependence tend to place high relevance on developing a *sense of belonging* than cultures that promote autonomy. Consequently, students from collectivistic cultures show high sensitivity to their peers' behaviors, recognize, and adopt well-discipline classmates' model behaviors, receive more positive feedback from the teachers, feel more successful at school and thus, have a high sense of belonging at school. Therefore, students' cultural background should be considered when evaluating sense of belonging at school (Chiu et al., 2016).

In this context, it is very likely that the sense of belonging at school scale is particularly sensitive to cultural-related features of the host countries which in turn impact the dynamics of educational systems within countries.

In terms of the MGCFA, the technique cannot handle all the features of ILSAs that should be considered in the evaluation of measurement invariance. For instance, the technique cannot handle many groups and it is not possible to evaluate differences within each country (e.g., native versus immigrant students). Thus, even though the findings did not provide evidence in favor of measurement invariance, the results cannot be conclusive due to the limitations of the technique and must be considered carefully.

Another aspect worth mentioning is the fact that this scale has both positively and negatively worded items and because of that, some items had to be reverse-coded. This is another issue that might have an impact in the psychometric properties of the instrument. As pointed by Kam and Fan (2020), positively and negatively worded items are supposed

to measure opposite sides of a construct but when used in the same scale, their correlation might be impacted which could lead to dimensionality issues where the instrument measures more than one construct: one related to the positively worded items and the other to the negatively worded. Furthermore, according to the authors previous research has shown that some test takers have difficulty responding to negatively worded items which can ultimately impact measurement invariance. Researchers and test developers are thus urged to evaluate the potential impact of the wording of the items and the extent to which it impacts the observed correlations among items and thus, the measurement of the latent construct before conducting evaluations of measurement invariance. The high number of flagged correlations between residuals could be then due also to the wording of the items.

The overall observed lack of measurement invariance of the sense of belonging at school scale could also be explained by some features of the scale there are some features of the scale that can explain this finding: (a) the cultural differences among countries in terms of the individualism/collectivism dimension that is directly related to the experience of sense of belonging at school, (b) the wording of the items in the scale that can have a potential impact in the response styles and could also lead to multidimensionality, and (c) the use of a Likert-type response scale. As noted by Weijters, Baumgartner, and Geuens (2016), the meaning of Likert-type response scales that include labels such as strongly (dis)agree or completely (dis)agree can systematically change across languages which in turn can lead to differences in the scale usage at the group level thus, response distributions can be different across groups due to the non-equivalency of the meaning of response categories. As a result, parameter estimates are

likely to be altered by cross-linguistic bias which in turn, decreases the likelihood of achieving measurement invariance.

The authors conducted a series of studies where they evaluated a coding method - the calibrated sigma method- to correct for group-level scale usage differences across groups of respondents from different countries. Their findings showed that the traditional Likert coding method led to biased results when administered to examinees from different countries in that results suggested an apparent difference in factor means across different language groups however, when they implemented the alternative method, measurement invariance testing did not lead to biased estimates of the latent means.

The evidence provided by these authors can help contextualize the findings from the MGCFA analysis in that students might be using the response scale in a different way due to language-related differences that can compromise the interpretation of the response options moreover, the finding suggests that the Likert-type response scale from the sense of belonging at school measure might not be suitable when administered to examines from diverse cultural backgrounds and thus, it highlights a methodological issue that should be controlled for from the test development stage. It is recommended that test developers collect sufficient empirical evidence about the psychometric functioning of the response scale and consider the implementation of alternate methods to code the responses that can reduce the likelihood of response style biases.

Finally, educators in general are urged to dedicate resources to better know the population of students they serve so that they can be aware of potential determinants of academic achievement and can identify the best ways to use that knowledge to inform their practices and policies. Educational administrators are also urged to be aware of the

limitations of the measurement instruments they use as input for the development of educational policies to avoid mistakes in the allocation of resources and more importantly, to avoid decisions that could be detrimental to the overall wellbeing of the students.

5.1.3 Reading Literacy

MGCFA was also implemented in the reading literacy scale and as mentioned, the analysis could only be performed on twelve countries due to convergence issues. It is interesting to note that the technique seems to have a different performance depending on whether the data are categorical or continuous since there were no convergence issues with the previous scales. This problem already pointed a limitation of the MGCFA when applied to data from ILSAs.

The overall findings from the MGCFA did not provide evidence in favor of the invariance of the measures. It is likely that despite the efforts made by PISA to guarantee the comparability of the reading measure, more research is needed to control for cultural factors that continue to impact test results.

In terms of the results, the configural model was just identified probably because it only included three indicators thus, no sound conclusions about the scale can be made based solely on the configural model. Results from the metric model did not show significant improvement over the just identified configural model moreover, the inspection of correlations between residuals showed problematic residuals in all the countries under analysis suggesting that the metric model was probably underpredicting the observed correlations.

This finding suggests that the latent construct could be expressed differently across the countries and thus, due to the diversity among test takers, it is likely that cultural differences could explain the lack of invariance. Ghorbandordinejad and Bayat (2014) have pointed that immigrant students have difficulties understanding the meaning of texts when they are not familiar with the culture from the host country which could in turn, impact their performance on reading comprehension tasks.

Reading comprehension is in fact a complex construct likely to be influenced by several factors (e.g., language, culture, economic status) that due to its nature is also sensitive to cultural differences thus, the achievement of measurement invariance can be challenging. As noted by Asil and Brown (2016), reading literacy is influenced by the features of a language system, the features of the writing system, approaches to teaching and learning that vary across cultures, and socioeconomic development all of which can explain measurement non-invariance. The authors also mention that the writing system has a major influence in reading performance as the reading tasks activate meaning and phonological systems of language.

In terms of the cultural differences, it is also known that cultural dimensions can impact both teaching and learning including the value and participation in high-stakes national and international testing. Some cultures prioritize testing and even provide additional training to students which can also introduce bias in ILSAs placing a threat in the validity of the interpretations from test scores. Asil and Brown (2016) mention six types of threats to the invariance of measures from PISA regarding the reading literacy test: language-specific differences in grammar, language-specific differences in writing, language-specific differences in meaning, cultural differences, translation strategies and

techniques, and editing-related problems. According to the authors there have also been some critics to the reading measure of PISA where it is thought that the reading passages favor Western countries. All together, these factors are likely to have impacted the item responses and could thus be potential sources of measurement non-invariance.

Regarding the scalar model, results pointed local fit problems and the evidence was not sufficient to establish measurement invariance suggesting that not only does the construct is manifested in different ways across countries but that test takers might also be approaching the response options differently. Typically, reading items have different response formats which can also introduce noise in the performance of students due to the same factors that were mentioned earlier and that tend to impact the overall performance in reading from students across different countries.

Qualitative methodologies are encouraged to address the non-equivalence problems and to identify ways to optimize the test contents by developing items and texts that are not sensitive to cultural differences. As with the previous scales, these findings are not conclusive due to the limitations of the modeling technique that was implemented to evaluate measurement invariance. Thus, it is possible that the findings were also impacted by the technique itself.

Based on these preliminary analyses, the MGCFA is not recommended as the only source to collect evidence about measurement invariance when the measures under analysis involve several groups and culturally diverse populations within each group. The use of this technique is likely to introduce methodological biases into the results compromising the validity of test interpretations. Researchers are encouraged to use alternate modeling techniques that are suitable to handle the complex features of data

from ILSAs and test users are also advised to be cautious when using data from ILSAs that do not include sound and sufficient evidence about measurement invariance. Moreover, policy makers are encouraged to use multiple sources of information -apart from the data from ILSAs- for decision-making purposes and make a responsible use of data from ILSAs by being aware of their limitations. It is important that test results are always interpreted in the context that is specific to each country and culture, and to the characteristics of the target student population.

5.2 Alignment Optimization

5.2.1 Exposure to Bullying Scale

The alignment optimization procedure was implemented next with the aim to overcome the limitations from MGCFA and collect more evidence about the extent to which the measures under analysis are invariant. One of the features of the alignment optimization procedure is that it can handle several groups at the same time making it possible to account for the immigration status of the students by dividing the countries into immigrants and native students as mentioned in the methodology section. The results for the exposure to bullying scale showed that the factor loadings of items 6 (“Other students spread nasty rumors about me”) and 3 (“I was threatened by other students”) were the most invariant across countries and students whereas items 4 (“Other students took away or destroyed things that belong to me”) and 5 (“I got hit or pushed around by other students”) were the least invariant. These two last items were designed to evaluate physical bullying thus, the findings could be suggesting that this specific dimension of bullying could probably have a different meaning both across countries and groups of students for instance, according to the results these two items were not invariant in Hong

Kong, Greece, Kazakhstan, Qatar, and United Arab Emirates. According to the cultural dimensions by Hosftede, these countries have similar scores in three dimensions: power distance, individualism/collectivism, and masculinity. Regarding power distance, the scores for these countries are high suggesting that in these societies people believe that inequalities are acceptable and more importantly, there is usually no defense against power abuse by superiors. In terms of individualism/collectivism, the countries scored low suggesting that the countries have a collectivist culture where people prioritize the interests of the group over themselves but mostly among family members because the relationships with non-family members can be hostile. Finally, regarding the masculinity dimension, the countries scored high suggesting that the societies are primarily masculine that is, driven by competition, achievement, and success.

This information provide context to explain the findings, it is likely that the lack of measurement invariance in these countries is in fact related to the cultural features that to some extent can impact the expression and experience of physical aggression. It is likely that the organization of the societies lead people to view these expressions as either acceptable or "normal".

Another interesting finding from this analysis was that the number of invariant factor loadings was consistently higher among immigrant students than among their native peers. This finding can be highlighting the fact that the immigrant students might be sharing a similar experience with respect to their status as immigrants. The situations that these students face could be very similar leading them to have a similar notion of the exposure to bullying, perhaps most of them have experienced bullying because of their

status as immigrants and if so, the results from the exposure to bullying scale are likely to lead to valid comparisons.

The results also provided evidence supporting metric invariance for items 6 ("Other students spread nasty rumors about me") and 3 ("I was threatened by other students") meaning that these items can ensure valid comparisons of the latent mean of bullying across countries and students. The findings from this technique are more informative and, in this sense, can guide the use of the test results by providing test users with information to make a responsible use of the results. Moreover, each country can easily identify to which other countries they can compare their test performance.

The analysis also provided information about the invariance of thresholds and the results indicated that items 5 ("I got hit or pushed around by other students") and 3 ("I was threatened by other students") showed the highest level of invariance both across countries and within students suggesting that students in general used the response categories in the same way thus, the information provided by these items can be compared across countries and students and are likely to lead to trustworthy comparisons. Items 1 ("Other students left me out of things on purpose") and 2 ("Other students made fun of me") on the other hand, were the least invariant across countries and students suggesting that the contents from these items could be likely to be interpreted in a different way depending on the country of residence and immigration status. Further inspection of the meaning of these items across countries and students are encouraged to identify the sources of invariance and make the necessary adjustments. As mentioned before, psychological constructs tend to be sensitive to cultural differences thus, it is possible that the features of some cultures influence the meaning of the statements

presented in the items. Information from these items must be used with caution and it is recommended to avoid international comparisons involving these items.

On the other hand, as with the previous finding, the thresholds are consistently more invariant across immigrant students and another interesting finding was that the United States was the only country where the thresholds were invariant across all the items and groups of students. This finding can provide feedback to educational policymakers as to the impact that their policies are having among the immigrant student population; it provides preliminary evidence suggesting that the educational policies could be promoting the proper inclusion of immigrant students into the educational systems.

In summary, the alignment optimization procedure provided more informative results as to the extent to which items are invariant across countries and students. By allowing the inclusion of groups of students within countries, it is easier to identify the countries and groups of students for which the items are invariant as defined by the alignment optimization method. The information provided by the analysis can help test users make responsible use of the information provided by the scale and it can also guide them as to know what countries and groups of students can be compared against one another increasing the trustworthiness of the measurement instrument and its suitability to be used for decision-making processes.

5.2.2 Sense of Belonging at School Scale

The alignment optimization procedure was implemented on the sense of belonging at school scale. In terms of the factor loadings results showed evidence suggesting that all the items from the scale were highly invariant across countries and

students and as with the previous analysis, the number of invariant factor loadings was consistently higher among immigrant students than across their native peers.

According to the results, there is empirical evidence suggesting that the sense of belonging at school construct is likely to be manifested in a similar way across countries and students thus, comparability of this construct among countries and students is likely to lead to valid inferences that can be used for decision-making processes. This finding, however, is not consistent with the findings from the MGCFA indicating the impact that the statistical modeling approach can have on the analysis of measurement invariance involving ILSAs.

Results also showed that the countries with the least number of invariant items included Qatar and Kazakhstan which is consistent with the findings from the bullying scale and given the cultural similarities between these countries, it was expected that they would have lowest levels of invariance. It would be important to conduct further inspection of the scale in these countries through the implementation of qualitative techniques such as cognitive interviews or focus groups to identify alternate ways of wording the items to reduce the impact of the culture and thus the biases related to it.

The analysis also showed that item 3 (“I feel like I belong at school”) showed the lowest number of invariant loadings among native students which could indicate real differences in the experience of sense of belonging at school between immigrant and native students. This finding could also be suggesting the need of more inclusive educational policies targeting immigrant students.

In terms of the invariance of item thresholds, the results showed a general lower number of invariant thresholds both across countries and students suggesting that even

though there is some level of invariance, it is likely that students across countries are using the response options in different ways. As mentioned before, the Likert-type scales tend to be sensitive to cultural differences that is, students from different cultures can interpret the response options differently introducing noise in the measurement process and increasing the likelihood of biased results. Moreover, this scale includes positive and negatively worded items which can also interfere with the interpretation of the response categories.

Based on the evidence provided by this analysis, it is recommended to avoid the use of negatively worded items and given the cultural diversity across countries, test developers are also advised to avoid the use of Likert-type scales and implement alternate response formats. Even though, the results do not provide evidence in favor of full measurement invariance, they did provide evidence suggesting that items 6 ("I feel lonely at school") and 4 ("I feel awkward and out of place in my school") are suitable indicators of sense of belonging at school across countries and students with different immigration status and thus, could be used for comparison purposes.

5.2.3 Reading Literacy

The alignment optimization procedure allowed for the inclusion of the 31 countries in the analysis as opposed to the MGCFA where the inclusion of the 31 countries led to convergence problems. Results showed that the subscale *understand* had the largest number of invariant factor loadings across countries and students whereas the subscale *locate information* showed the lowest number. Overall, the results indicated that the indicators of the reading literacy scale are likely to be interpreted differently across countries and students. This finding is consistent with that from the MGCFA and it could

be providing evidence about the extent to which the cultural background of the test takers might be interfering with the measurement of reading.

The results related to the intercepts showed a similar scenario where the intercepts tend to vary across countries and students suggesting that students are using the response categories differently and thus, might be interpreting it in different ways. The overall results, suggest that the meaning of the reading literacy indicators does not hold across countries and students.

These findings are aligned with a well-documented finding from educational research targeting immigrant students is the presence of an educational achievement gap between immigrant and native students where the former shows lower academic performance, especially in reading and writing measures (Arikan et al., 2017; Azzolini et al., 2012; Giannelli & Rapallini, 2016; Powers & Pivovarova, 2017; Teltemann & Schunck, 2016).

Test developers are urged to conduct an ongoing evaluation of the reading literacy items while promoting cross-cultural research studies that could provide information about alternate item and response formats that could help to reduce cultural-related biases from the reading measures. Educational policymakers and educators in general are advised to make careful use of the findings from this measure and be aware of its limitations when using the results for decision-making processes. Educational institutions should not make inferences about the students based solely on the results from this assessment given the findings from these analyses.

In summary, the alignment optimization procedure is a suitable alternative to evaluate measurement invariance on data from ILSAs. The technique allows for the

inclusion of several groups which in turns, allows for the evaluation of subgroups within each country. Moreover, the alignment optimization provides more informative and detailed results that can help test users to make a responsible use of the results. This statistical tool identifies the extent to which the measures are invariant even if full invariance is not achieved increasing the possibilities to make proper use of the test scores.

5.3 Evaluation of the Relationship between the Non-cognitive Measures and the Performance on Reading Literacy

Another aim of this dissertation was to identify the extent to which the non-cognitive measures could be impacting the performance on the reading literacy scale and to do so, a latent structural regression model was evaluated. Results from this analysis provided evidence suggesting that the two non-cognitive constructs (exposure to bullying and sense of belonging at school) were statistically significant predictors of reading literacy so that high levels of sense of belonging at school were associated to higher performance in the reading literacy scale whereas higher levels of exposure to bullying were associated to lower performance.

However, the findings from this model should be interpreted in the context of the analyses of measurement invariance. Even though, the direction of the relationships is consistent to what it would be expected theoretically, the fact that the measures are not fully invariant and the particularly high lack of invariance of the reading literacy measure is likely to have influenced the estimations from the latent structural regression model and thus, the results are likely to be biased in that the model might not be reflecting the true underlying relationships among the constructs.

The evaluation of measurement invariance should be the starting point to conduct further analysis and it should be used as a context for the interpretation of the results. Given the educational needs of governments around the world, the tendency is to use results from ILSAs to conduct several forms of predictive modeling to identify predictors that can be manipulated to increase academic performance and use it as an indicator of the quality of educational systems.

The structural latent regression model tested in this dissertation could seem as providing useful information that could potentially be used for the formulation of educational policies however, given the lack of invariance of the measures, these results should not be directly used for decision-making purposes and should be evaluated with caution.

CHAPTER

VI CONCLUSION

This dissertation aimed to provide evidence about the extent to which the 2018 Programme for International Student Assessment (PISA) provided invariant measures of reading literacy, exposure to bullying, and sense of belonging at school for immigrant students from diverse cultural and linguistic backgrounds across the countries that host large populations of immigrants.

The decision to conduct this research was based on the large increase of migration movements throughout the world and the need to provide governments with information they can use to integrate immigrant students into their educational systems while ensuring the high quality of the education and the academic success of the students.

Migration movements have become a great concern among governments around the world in that they represent several threats to the social cohesion and economy of the countries however, one way to address the issue is through the proper integration of immigrants into the social systems of the host countries so that they can contribute to the overall development of the society. To do so, the first step is to provide immigrants with the resources they need to develop skills so that they can have an active role in the society and thus, can contribute to its development.

In this scenario, educational systems play a key role in the proper integration of immigrants into the social system by helping them to develop all the skills they need to become competent and active members of the society. However, the success of the educational processes and practices depends largely on the proper assessment systems that provide institutions with accurate information about (a) the academic skills of the

students, (b) the impact of the educational policies and practices, and (c) the areas within the educational system that need to be improved or modified.

PISA is one of the most widely used educational measures that provides countries with a standard against which they can judge the overall quality of their educational systems and in this context, PISA can be a powerful tool in that it can be used to modify/improve educational policies throughout the world. However, the potential impact of this assessment can be hindered by the complexity of the target population of students which is becoming more and more diverse as the migration movements continue to increase. Thus, in order to make proper use of the assessment instruments and its results, evidence about the extent to which those instruments provide measures of cognitive and non-cognitive constructs regardless of the cultural background or immigration status of the students is needed.

This dissertation aimed to provide this evidence in an effort to promote the responsible use of international large-scale assessments so that countries throughout the world can enhance the quality of their educational systems while promoting social cohesion in the light of the challenges proper of this decade specifically, migration movements.

The overall findings have implications for educational policymakers, educators, test developers, and educational researchers. Policymakers are urged to make a responsible use of the results from ILSAs by avoiding making decisions based only on these assessments, collecting multiple sources of data that can inform their decision-making processes, be aware of the impact that the cultural background of the students can have on their overall academic performance and reinforce inclusive educational practices

to alleviate this impact, be cautious when trying to emulate educational practices from countries with outstanding test performance by tailoring the practices to the characteristics and needs of their student population, promoting/financing evidence-based research to better characterize the immigrant student population with the aim to properly identify their most urgent needs, and invest resources in the training of teachers to provide them with the additional tools they might need to respond to the immigrant students' needs.

Policymakers could also make efforts to identify best inclusion practices implemented in other countries to improve their current policies. In this dissertation, it was interesting to see in the MGCFA analysis of the bullying scale that Canada was among the countries that contributed to the chi-square because as noted by Ardakani et al. (2011) Canada is one of the countries that has made more efforts towards the internationalization of the curriculum. In fact, most educational institutions supervise the experiences of teachers with the aim to promote intercultural international learning. In this context, the fact that Canada was among the countries that deteriorated the Chi-square could be reflecting a difference in terms of the way this country approaches intercultural learning thus, in the case of Canada the finding could be highlighting a different, but positive educational approach that could be beneficial for immigrant students.

On the other hand, educators are similarly urged to educate themselves to become familiar with the cultural backgrounds of the immigrant students so that they can better tailor the educational materials and resources to their needs, be aware of the limitations of ILSAs when analyzing and using the results, be aware that the limitations of ILSAs can

also apply to their own assessment tools, avoid the tendency to assume that immigrant students are interpreting the educational materials in the same way as their peers, be willing to providing accommodations when needed, and promote an inclusive climate in their classrooms.

Test developers are urged to implement qualitative research techniques prior to the development of test items to collect as much information as possible regarding the cultural practices of the target population of test takers to prevent cultural biases, explore and implement different item and response formats to control for response style-related biases, when developing items for psychological-related measures avoid language that could trigger stress responses, implement different translation techniques, and be always aware of the impact that cultural systems can have in every section of the measurement instrument: (a) cultural systems determine the meaning and characteristics of cognitive processes and psychological constructs, (b) methodological biases such as translation biases, perception of response styles, lack of familiarity with testing procedures, and construct underrepresentation can affect specific items or the whole instrument violating the requirements for measurement equivalence, and (c) the lack of generalizability of the constructs from the individual level to the national or cultural level can impact the observed differences on test scores (Samara et al., 2019).

Educational researchers are encouraged to conduct measurement invariance analysis prior to using data from ILSAs to identify the extent to which they can draw accurate interpretations that are based on comparisons across countries. Specifically, qualitative research techniques could be implemented to identify how students living in these countries are interpreting the test items and the response scale. Similarly,

governments throughout the world should take precautions when using the results from ILSAs for decision-making processes; it is necessary that they do not rely in the results from these assessments only to draw conclusions about the quality of their educational systems but perhaps more importantly, they are encouraged to promote evidence-based research to provide a context in which test results can be properly interpreted so that they can be used as the basis for the improvement and development of educational policies that increase the overall quality of education for all the student population they serve.

Educational researchers are encouraged to promote research studies to evaluate the scope and limitations of the most commonly used techniques to assess measurement invariance, promote the implementation of more sophisticated statistical modeling techniques that can account for the complexities inherent to ILSAs (the nesting of students within countries, the sampling procedures, the large amounts of missing data by test design, large number of groups under analysis) such as multilevel structural equation modeling, and continue to collect empirical evidence about the performance of the alignment optimization procedure when applied to data from ILSAs.

Finally, there were some limitations of this study that should be addressed in future studies: (a) data from ILSAs usually involves large amounts of missing data by design and the analyses conducted in this dissertation did not include any techniques to handle the missing data, missing data was delete it which leads to loss of information, (b) the nesting of the students within the countries was not accounted for by the techniques implemented to evaluate measurement invariance, (c) the reading literacy scale included different features than the non-cognitive scales such as: less number of countries under analysis when the MGCFA was implemented due to the convergence issues, the analyses

were conducted at the test level instead of the item level, and the measurement model only included three indicators; all of these issues could have impacted the observed results and thus, the findings for this scale are not conclusive and must be interpreted with caution, (d) the countries to be analyzed were not selected randomly but according to the sample size which can introduce methodological bias in the overall findings, (e) the generalizability of the findings is limited due to these limitations, and (f) the countries that contributed the most to the Chi Square were kept in the analyses however, further analyses on these data could remove those countries and evaluate model performance.

Future studies should be implemented to address these limitations and continue to collect empirical evidence as the extent to which the alignment optimization procedure is a suitable technique to evaluate the invariance of cognitive and non-cognitive scales from PISA.

APPENDICES

APPENDIX A
MPLUS CODE MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS BULLYING/SENSE OF BELONGING AT SCHOOL SCALE

```
TITLE:    MGCFA BULLYING Configural invariance
  Data:
  file is "\bullyingMGCFA.dat";
variable: names are COUNTRY x1-x6 IMM;
        grouping is COUNTRY (36 = Australia 40 = Austria 56 = Belgium
96 = Brunei 124 = Canada 188 = CostaR 191 = Croatia 208 = Denmark
233 = Estonia 250 = France 276 = Germany 300 = Greece 344 = HongK
372 = Ireland 380 = Italy 398 = Kazak 442 = Luxem 446 = Macao
528 = Nether 554 = NewZ 578 = Norway 634 = Qatar 688 = Serbia
702 = Singap 705 = Slovenia 724 = Spain 752 = Sweden 756 = Switz
784 = UArab 826 = UK 840 = US);
        categorical are x1-x6;
        usevariables x1-x6;
analysis: parameterization = theta; estimator = wlsmv;

model: bullying by x1@1 x2 x3 x4 x5 x6;
          x1-x6@1;  !Fix residual variances to 1

model Austria:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model Belgium:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model Brunei:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model Canada:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model CostaR:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model Croatia:
bullying by x1@1 x2 x3 x4 x5 x6;
          [x2$1-x6$3];

model Denmark:
```

```
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Estonia:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model France:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Germany:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Greece:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model HongK:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Ireland:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Italy:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Kazak:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Luxem:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Macao:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Nether:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model NewZ:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];
```

```
model Norway:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Qatar:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Serbia:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Singap:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Slovenia:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Spain:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Sweden:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model Switz:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model UArab:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model UK:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];

model US:
bullying by x1@1 x2 x3 x4 x5 x6;
        [x2$1-x6$3];
output: sampstat residual tech1 stdyx modindices(4.00);
savedata: difftest="C:\ dif.diff";
```

```
TITLE:    MGCFA BULLYING Metric invariance
  Data:
  file is "C: \bullyingMGCFA.dat";
variable: names are COUNTRY x1-x6 IMM;
        grouping is COUNTRY (36 = Australia 40 = Austria
56 = Belgium 96 = Brunei 124 = Canada 188 = CostaR
191 = Croatia 208 = Denmark 233 = Estonia 250 = France
276 = Germany 300 = Greece 344 = HongK 372 = Ireland
380 = Italy 398 = Kazak 442 = Luxem 446 = Macao
528 = Nether 554 = NewZ 578 = Norway 634 = Qatar 688 = Serbia
702 = Singap 705 = Slovenia 724 = Spain 752 = Sweden 756 = Switz
784 = UArab 826 = UK 840 = US);
        categorical are x1-x6;
        usevariables x1-x6;
analysis: parameterization = theta; estimator = wlsmv;
difftest="C:\ dif.diff";
model: bullying by x1@1 x2 x3 x4 x5 x6;
          x1-x6@1;
model Austria:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model Belgium:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model Brunei:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model Canada:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model CostaR:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model Croatia:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
          [x6$2];[x6$3];

model Denmark:
          [x2$2];[x2$3];[x3$2];[x3$3];
          [x4$2];[x4$3];[x5$2];[x5$3];
```

```
        [x6$2];[x6$3];

model Estonia:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model France:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Germany:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Greece:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model HongK:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Ireland:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Italy:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Kazak:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Luxem:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model Macao:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];
```

model Nether:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model NewZ:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Norway:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Qatar:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Serbia:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Singap:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Slovenia:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Spain:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Sweden:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

model Switz:
[x2$2];[x2$3];[x3$2];[x3$3];
[x4$2];[x4$3];[x5$2];[x5$3];
[x6$2];[x6$3];

```
model UArab:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model UK:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];

model US:
        [x2$2];[x2$3];[x3$2];[x3$3];
        [x4$2];[x4$3];[x5$2];[x5$3];
        [x6$2];[x6$3];
 output: sampstat residual tech1 stdyx modindices(4.00);
savedata: difftest="C:\difmetricfbull.dat";
```

```
TITLE:    MGCFA BULLYING Scalar invariance
  Data:
  file is "C:\ bullyingMGCFA.dat";
variable: names are COUNTRY x1-x6 IMM;
        grouping is COUNTRY (36 = Australia 40 = Austria
56 = Belgium 96 = Brunei 124 = Canada 188 = CostaR
191 = Croatia 208 = Denmark 233 = Estonia 250 = France
276 = Germany 300 = Greece 344 = HongK 372 = Ireland
380 = Italy 398 = Kazak 442 = Luxem 446 = Macao
528 = Nether 554 = NewZ 578 = Norway 634 = Qatar 688 = Serbia
702 = Singap 705 = Slovenia 724 = Spain 752 = Sweden 756 = Switz
784 = UArab 826 = UK 840 = US);
        categorical are x1-x6;
        usevariables x1-x6;
analysis: parameterization = theta; estimator = wlsmv;
difftest is "C:\difmetricfbull.dat";
model: bullying by x1@1 x2 x3 x4 x5 x6;
          x1-x6@1;  !Fix residual variances to 1
bullying*;
[bullying@0];


output: sampstat residual tech1 stdyx modindices(4.00);
```

APPENDIX B

MPLUS CODE MULTIPLE GROUP CONFIRMATORY FACTOR ANALYSIS READING LITERACY SCALE

```
TITLE:    MGCFA READING Configural invariance

Data:
  file is "C:\ reading.dat";

variable: names are COUNTRY x1-x3 IMM;
USEOBSERVATIONS = COUNTRY LE 300;
        grouping is COUNTRY (36 = Australia 40 = Austria 56 = Belgium
96 = Brunei 124 = Canada 188 = CostaR 191 = Croatia 208 = Denmark
233= Estonia 250 = France 276 = Germany 300 = Greece);


usevariables x1-x3;

analysis: type= general; estimator = ML;

model: reading by x1@1 x2 x3; !Default loading of first item fixed to 1.


model Austria:
reading by x1@1 x2 x3;
          [x2-x3];

model Belgium:
reading by x1@1 x2 x3;
          [x2-x3];

model Brunei:
reading by x1@1 x2 x3;
          [x2-x3];

model Canada:
reading by x1@1 x2 x3;
          [x2-x3];

model CostaR:
reading by x1@1 x2 x3;
          [x2-x3];

model Croatia:
reading by x1@1 x2 x3;
          [x2-x3];

model Denmark:
reading by x1@1 x2 x3;
```

```
        [x2-x3];

model Estonia:
reading by x1@1 x2 x3;
        [x2-x3];

model France:
reading by x1@1 x2 x3;
        [x2-x3];

model Germany:
reading by x1@1 x2 x3;
        [x2-x3];

model Greece:
reading by x1@1 x2 x3;
        [x2-x3];


output: tech1 sampstat residual stdyx;
```

```
TITLE:    MGCFA READING Metric invariance

Data:
  file is "C:\ reading.dat";

variable: names are COUNTRY x1-x3 IMM;
        grouping is COUNTRY (36 = Australia 40 = Austria 56 = Belgium
96 = Brunei 124 = Canada 188 = CostaR 191 = Croatia 208 = Denmark
233= Estonia 250 = France 276 = Germany 300 = Greece);


usevariables x1-x3;

analysis: type= general; estimator = ML;

model: reading by x1@1 x2 x3;


model Austria:
reading
          [x2-x3];

model Belgium:
reading
          [x2-x3];

model Brunei:
reading
          [x2-x3];

model Canada:
reading
          [x2-x3];

model CostaR:
reading
          [x2-x3];

model Croatia:
reading
          [x2-x3];

model Denmark:
reading
          [x2-x3];

model Estonia:
reading
          [x2-x3];

model France:
```

```
reading
        [x2-x3];

model Germany:
reading
        [x2-x3];

model Greece:
reading
        [x2-x3];


output: tech1 sampstat residual stdyx;
```

APPENDIX C

MPLUS CODE ALIGNMENT OPTIMIZATION BULLYING/BELONG/READING LITERACY SCALE

```
DATA: FILE ="C:\bullying.dat";
VARIABLE:
NAMES = COUNTRY label SCHOOLID Y1 Y2 Y3 Y4 Y5 Y6 IMM W_FSTUWT newimm;
USEVARIABLES= Y1 Y2 Y3 Y4 Y5 Y6 CONTSCHL;
CLUSTER=CONTSCHL;
WEIGHT = W_FSTUWT;
classes = c(72);
knownclass = c(COUNTRY=32 3200 36 3600 40 4000 56 5600 96 9600
124 12400 188 18800 191 19100 208 20800 233 23300 250 25000
276 27600 300 30000 344 34400 372 37200 380 38000 398 39800
400 40000 442 44200 446 44600 499 49900 528 52800 554 55400
578 57800 634 63400 643 64300 682 68200 688 68800 702 70200
705 70500 724 72400 752 75200 756 75600 784 78400 826 82600
840 84000);
DEFINE: CONTSCHL= (COUNTRY*10000)+ SCHOOLID;
ANALYSIS:
TYPE=MIXTURE COMPLEX;
ESTIMATOR=MLR;
ALIGNMENT= FIXED(52800);
PROCESSORS=6;
MODEL:
%OVERALL%
bullying BY Y1 Y2 Y3 Y4 Y5 Y6;
OUTPUT:
tech1 tech8 align SVALUES;
```

References

Akresh, R., & Akresh, I. R. (2011). Using achievement tests to measure language assimilation and language bias among the children of immigrants. *The Journal of Human Resources, 46*(3), 647- 667. http://doi.org/10.3368/jhr.46.3.647

Alivernini, F., Manganelli, S., Cavicchiolo, E., & Lucidi, F. (2019). Measuring bullying and victimization among immigrant and native primary school students: evidence from Italy. *Journal of Psychoeducational Assessment, 37*(2), 226-238. http://doi.org/ 10.1177/0734282917732890

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington: American Educational Research Association.

Ardakani, F. B., Yarmohammadian, M. H., Abari, A. A. F., & Fathi, K. (2011). Internationalization of higher education systems. *Procedia Social and Behavioral Sciences, 15*, 1690-1695. https://doi.org/10.1016/j.sbspro.2011.03.353

Areepattamannil, S., & Kaur, B. (2012). Factors predicting science achievement of immigrant and non-immigrant students: a multilevel analysis. *International Journal of Science and Mathematics Education, 11*, 1183-1207. https://doi.org/10.1007/s10763-012-9369-5

Arikan, S., van de Vijver, F. J. R., & Yagmur, K. (2017). PISA mathematics and Reading performance differences of mainstream European and Turkish immigrant students. *Educational Assessment, Evaluation and Accountability, 29*, 229-246. http://doi.org/10.1007/s11092-017-9260-6

Asil, M., & Brown, G. T. L. (2016). Comparing OECD PISA Reading in English to other languages: identifying potential sources of non-invariance. *International Journal of Testing, 16(1)*, 71-93. https://doi.org/10.1080/15305058.2015.1064431

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397-438. http://doi.org.10.1080/10705510903008204

Asparouhov, T., & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 1-14. http://doi.org/10.1080/10705511.2014.919210

Avvisati, F. (2020). The measure of socio-economic status in PISA: a review and some suggested improvements. *Large-scale Assessments in Education. An IEA-ETS Research Institute Journal, 8*, 1-37. https://doi.org/10.1186/s40536-020-00086-x

Azzolini, D., Schnell, P., & Palmer, J. R. B. (2012). Educational achievement gaps between immigrant and native students in two "new" immigration countries: Italy and Spain in comparison. *The ANNALS of the American Academy of Political and Social Science, 643*(1), 46-77. http://doi.org/10.1177/0002716212441590

Borgonovi, F. (2018). *"How do the performance and well-being of students with an immigrant background compare across countries?"* (PISA in Focus, No. 82). OECD Publishing. https://doi.org/10.1787/a9e8c1ab-en

Bozick, R., Malchiodi, A., & Miller, T. (2016). Premigration school quality, time spent in the United States, and the math achievement of immigrant high school students. *Demography, 53*, 1477-1498. http://doi.org/10.1007/s13524-016-0497-3

Braun, H., & von Davier, M. (2017). The use of test scores from large-scale assessment surveys: psychometric and statistical considerations. *Large-scale Assessments in Education, 5*(17), 1-16. http://doi.org/10.1186/s40536-017-0050-x

Byrne, B. M. (2004). Testing for multigroup invariance using AMOS graphics: a road less traveled. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(2), 271-300. http://doi.org/10.1207/s15328007sem1102_8

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: implications for improved education and training in Psychology. *Training and Education in Professional Psychology, 3*(2), 94-105. http://doi.org/10.1037/a0014516

Byrne, B. M., & van de Vijver F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: addressing the issue of nonequivalence. *International Journal of Testing, 10*, 107-132. http://doi.org/10.1080/15305051003637306

Byrne, B. M., & van de Vijver, F. J. R. (2017). The maximum likelihood alignment approach to testing for approximate measurement invariance: a paradigmatic cross-cultural application. *Psicothema, 29*(4), 539-551. http://doi.org/10.7334/psicothema2017.178

Callahan, R., Wilkinson, L., & Muller, C. (2010). Academic achievement and course taking among language minority youth in U.S. schools: effects of ESL placement. *Educational Evaluation and Policy Analysis, 32*(1), 84–117. http://doi.org/10.3102/0162373709359805

Carter, N. T., Kotrba, L. M., & Lake, C. J. (2014). Null results in assessing survey score comparability: illustrating measurement invariance using item response theory.

Journal of Business and Psychology, 29(2), 205-220. http://doi.org/10.1007/s10869-012-9283-4

Casper, D. M., Meter, D. J., & Card, N. A. (2015). Addressing measurement issues related to bullying involvement. *School Psychology Review, 44*(4), 353-371. http://doi.org/10.17105/spr-15-0036.1

Cattaneo, M. A., & Wolter, S. C. (2015). Better migrants, better PISA results: findings from a natural experiment. *IZA Journal of Migration, 4*, 1-19. http://doi.org/10.1186/s40176-015-0042-y

Cheung, M. W. L., & Au, K. (2005). Applications of multilevel structural equation modeling to cross-cultural research. *Structural Equation Modeling, 12*(4), 598-619. http://doi.org/10.1207/s15328007sem1204_5

Cheung, M. W. L., Leung, K., & Au, K. (2006). Evaluating multilevel models in cross-cultural research. An illustration with social axioms. *Journal of Cross-cultural Psychology, 37*(5), 522-541. http://doi.org/10.1177/0022022106290476

Chiu, M. M., Chow, B. W., McBride, C., & Mol, S. T. (2016). Students' sense of belonging at school in 41 countries: cross-cultural variability. *Journal of Cross-Cultural Psychology, 47*(2), 175-196. http://doi.org/10.1177/0022022115617031

Chiu, M. M., Pong, S., Mori, I., & Chow, B. W. (2012). Immigrant students´ emotional and cognitive engagement at school: a multilevel analysis of students in 41 countries. *Journal of Youth and Adolescence, 41*(11), 1409-1425. http://doi.org/10.1007/s10964-012-9763-x

Christ, O., Hewstone, M., Schmid, K., Green, E. G. T., Sarrasin, O., Gollwitzer, M., & Wagner, U. (2017). Advanced multilevel modeling for a science of groups: a short primer on multilevel structural equation modeling. *Group Dynamics: Theory, Research, and Practice, 21*(3), 121-134. http://doi.org/10.1037/gdn0000065

Cieciuch, J., Davidov, E., Schmidt, P., Algesheimer, R., & Schwartz, S. H. (2014). Comparing results of an exact vs. an approximate (Bayesian) measurement invariance tests: a cross-country illustration with a scale to measure 19 human values. *Frontiers in Psychology, 5*, 1-10. http://doi.org/10.3389/fpsyg.2014.00982

Cordero, J. M., Cristóbal, V., & Santín, D. (2018). Causal inference on education policies: a survey of empirical studies using PISA, TIMSS and PIRLS. *Journal of Economic Surveys, 32*(3), 878-915. http://doi.org/10.1111/joes.12217

Craig, W., Harel-Fisch, Y., Fogel-Grinvald, H., Dostaler, S., Hetland, J., Simons-Morton, B., … HBSC Bullying Writing Group. (2009). A cross-national profile of bullying and victimization among adolescents in 40 countries. *International*

Journal of Public Health, 54(2), 216-224. http://doi.org/10.1007/s00038-009-5413-9

Crosnoe, R., & Turley, R. N. (2011). K-12 educational outcomes of immigrant youth. *The Future of Children, 21*(1), 129-152. http://doi.org/10.1353/foc.2011.0008

Davidov, E., Dülmer, H., Cieciuch, J., Kuntz, A., Seddig, D., & Schmidt, P. (2018). Explaining measurement nonequivalence using multilevel structural equation modeling: the case of attitudes toward citizenship rights. *Sociological Methods & Research, 47*(4), 729-760. http://doi.org/10.1177/0049124116672678

Desouky, T. F., Mora, P. A., & Howell, E. A. (2013). Measurement invariance of the SF-12 across European-American, Latina, and African American postpartum women. *Quality of Life Research, 22*(5), 1135-1144. http://doi.org/10.1007/s11136-012-0232-5

Drasgow, F., & Probst, T. M. (2005). The psychometrics of adaptation: evaluating measurement equivalence across languages and cultures. In R. K. Hambleton, P.F. Merenda & C. D. Spielberg (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 265-294). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Dunn, E. C., Masyn, K. E., Jones, S. M., Subramanian, S. V., & Koenen, K. C. (2015). Measuring psychosocial environments using individual responses: an application of multilevel factor analysis to examining students in schools. *Prevention Science, 16*(5), 718-733. http://doi.org/10.1007/s11121-014-0523-x

Duong, M. T., Badaly, D., Liu, F. F., Schwartz, D., & McCarty, C. A. (2016). Generational differences in academic achievement among immigrant youths: a meta-analytic review. *Review of Educational Research, 86*(1), 3-41. http://doi.org/10.3102/0034654315577680

Educational Testing Service. (2014). *ETS Standards for Quality and Fairness*. https://www.ets.org/s/about/pdf/standards.pdf

Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in Psychology* (pp. 179-215). New York, NY: Cambridge University Press.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-group invariance testing possibilities in R. *Frontiers in Psychology*, *10* (1507), 1-18. http://doi.org/10.3389/fpsyg.2019.01507

Flouri, E., & Papachristou, E. (2019). Peer problems, bullying involvement, and affective decision-making in adolescence. *British Journal of Developmental Psychology, 37*(4), 466-485. http://doi.org/10.1111/bjdp.12287

Ghorbandordinejad, F., & Bayat, Z. (2014). The effect of cross-cultural background knowledge instruction on Iranian EFL learners' reading comprehension ability. *Theory and Practice in Language Studies*, *4*(11), 2373-2383. http://doi.org/10.4304/tpls.4.11.2373-2383

Giannelli, G. C., & Rapallini, C. (2016). Immigrant student performance in Math: does it matter where you come from? *Economics of Education Review*, *52*, 291-304. http://doi.org/10.1016/j.econedurev.2016.03.006

Global Migration Data Analysis Centre & International Organization for Migration. (2018). *Global Migration Indicators 2018*. Global Migration Data Analysis Centre. https://publications.iom.int/system/files/pdf/global_migration_indicators_2018.pdf

Gorges, J., Koch, T., Maehler, D. B., & Offerhaus, J. (2017). Same but different? Measurement invariance of the PIAAC motivation-to-learn scale across key socio-demographic groups. *Large-scale Assessments in Education*, *5*(13), 1-28. http://doi.org/10.1186/s40536-017-0047-5

Green, E. G. T., Deschamps, J., & Páez, D. (2005). Variation of individualism and collectivism within and between 20 countries. A typological analysis. *Journal of Cross-cultural Psychology*, *36*(3), 321-339. http://doi.org/10.1177/0022022104273654

Halamová, J., Kanovský, M., Gilbert, P., Troop, N.A., Zuroff, D. C., Petrocchi, N., Hermanto, N., Krieger, T., Kirby, J. N., Asano, K., Matos, M., Yu, F., Sommers-Spijkerman, M., Shahar, B., Basran, J., & Kupeli, N. (2019). Multiple group IRT measurement invariance analysis of the forms of self-criticising/attacking and self-reassuring scale in thirteen international samples. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *37*, 411-444. https://doi.org/10.1007/s10942-019-00319-1

He, J., Barrera-Pedemonte, F., & Buchholz, J. (2019). Cross-cultural comparability of noncognitive constructs in TIMSS and PISA. *Assessment in Education: Principles, Policy & Practice*, *26*(4), 369-385. http://doi.org/10.1080/0969594X.2018.1469467

Hofstede, G. (2011). Dimensionalizing Cultures: The Hofstede Model in Context. *Online Readings in Psychology and Culture*, *2*(1). https://doi.org/10.9707/2307-0919.101

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of*

Educational Research, 62(3), 333-353. http://doi.org/10.1080/00313831.2016.1258726

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods, 6*(2), 87-93. http://doi.org/10.18148/srm/2012.v6i2.5033

Hussein, M. H. (2010). The peer interaction in primary school questionnaire: testing for measurement equivalence and latent mean differences in bullying between gender in Egypt, Saudi Arabia and the USA. *Social Psychology of Education, 13*, 57-76. http://doi.org/10.1007/s11218-009-9098-yInternational Test Commission, (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf

International Test Commission. (2017). *ITC Guidelines for translating and adapting tests (second edition)*. guideline_test_adaptation_2ed.pdf (intestcom.org)

International Test Commission. (2018). *ITC Guidelines for the large-scale assessment of linguistically and culturally diverse populations*. https://www.intestcom.org/files/guideline_diverse_populations.pdf

Isac, M. M., Palmerio, L., & van der Werf, M. P. C. (2019). Indicators of (in)tolerance toward immigrants among European youth: an assessment of measurement invariance in ICCS 2016. *Large-scale Assessments in Education, 7*(6), 1-21. http://doi.org/10.1186/s40536-019-0074-5

Jak, S. (2014). Testing strong factorial invariance using three-level structural equation modeling. *Frontiers in Psychology, 5*, 1-7. https://doi.org/10.3389/fpsyg.2014.00745

Jak, S., Oort, F. J., & Dolan, C. V. (2014). Measurement bias in multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal, 21*(1), 31-39. http://doi.org/10.1080/10705511.2014.856694

Kam, C. C. S., & Fan, X. (2020). Investigating response heterogeneity in the context of positively and negatively worded items by using factor mixture modeling. *Organizational Research Methods, 23*(2), 322-341. https://doi.org/10.1177/1094428118790371

Kaplan, D. (1995). The impact of BIB spiraling. Induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Sciences, 20*(1), 69-82. https://doi.org/10.2307/1165388

Kim, E. S., Cao, C., Wang, Y., & Nguyen, D. T. (2017). Measurement invariance testing with many groups: a comparison of five approaches. *Structural Equation*

Modeling: A Multidisciplinary Journal, 24, 524-544. http://doi.org/10.1080/10705511.2017.1304822

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*(6), 881-898. http://doi.org/10.1080/00273171.2016.1228042

Kline, R. B. (2016). *Principles and practice of structural equation modeling*. New York, NY: The Guilford Press

Lamm, R., Do, T., & Rodriguez, M. C. (2019, April). *Measurement Invariance of An International Developmental Assets Measure: Alignment of 29 Countries* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, Toronto, Canada.

Lee, S., Bulut, O., & Suh, Y. (2017). Multidimensional extension of multiple indicators multiple causes models to detect DIF. *Educational and Psychological Measurement, 77*(4), 545-569. http://doi.org/10.1177/0013164416651116

Lee, J., Little, T. D., & Preacher, K. J. (2011). Methodological issues in using structural equation models for testing differential item functioning. In E. Davidov, P. Schmidt, J. Billiet, & B. Meuleman (Eds.), *Cross-cultural Analysis Methods and Applications* (pp. 55-84). New York, NY: Taylor and Francis Group.

Lomazzi, V. (2018). Using alignment optimization to test the measurement invariance of gender role attitudes in 59 countries. *Methods, data, analyses, 12*(1), 77-104. http://doi.org/10.12758/mda.2017.09

Marsh, H. W., Guo, J., Parker, P. D., Nagengast, B., Asparouhov, T., Muthén, B., & Dicke, T. (2018). What to do when scalar invariance fails: the extended alignment method for multi-group factor analysis comparison of latent means across many groups. *Psychological Methods, 23*(3), 524-545. doi: 10.1037/met0000113

Marsh, H. W., Nagengast, B., Morin, A. J. S., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: an application of exploratory structural equation modeling. *Journal of Educational Psychology, 103*(3), 701-732. http://doi.org/10.1037/a0024122

Martin, A. J., Liem, G. A. D., Mok, M. M. C., & Xu, J. (2012). Problem solving and immigrant student mathematics and science achievement: multination findings from the Programme for International Student Assessment (PISA). *Journal of Educational Psychology, 104*(4), 1054-1073. http://doi.org/10.1037/a0029152

Martin, S. R., Williams, D. R., & Rast, P. (2019). Measurement invariance assessment with Bayesian hierarchical including modeling. *PsyArXiv,* 1-16. http://doi.org/10.31234/osf.io/qbdjt

Melendez-Torres, G. J., Hewitt, G., Hallingberg, B., Anthony, R., Collishaw, S., Hall, J., Murphu, S., & Moore, G. (2019). Measurement invariance properties and external construct validity of the short Warwick-Edinburgh mental wellbeing scale in a large national sample of secondary school students in Wales. *Health and Quality of Life Outcomes, 17*, 1-9. http://doi.org/10.1186/s12955-019-1204-z

Meng, L., Qiu, C., & Boyd-Wilson, B. (2018). Measurement invariance of the ICT engagement construct and its association with students' performance in China and Germany: Evidence from PISA 2015 data. *British Journal of Educational Technology,0*, 1-19. http://doi.org/10.1111/bjet.12729

Meuleman, B. (2019). Multilevel structural equation modeling for cross-national comparative research. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie, 71*(1), 129-155. http://doi.org/10.1007/s11577-019-00605-x

Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. http://doi.org/10.1007/s11336-007-9039-7

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3), 479-515. http://doi.org/10.1207/S15327906MBR3903_4

Mplus Discussion. (2018, March 6). *Alignment Method Question* [Online forum post]. Mplus. http://www.statmodel.com/discussion/messages/9/24842.html?1520389172

Munck, I., Barber, C., & Torney-Purta, J. (2018). Measurement invariance in comparing attitudes toward immigrants among youth across Europe in 1999 and 2009: the alignment method applied to IEA CIVED and ICCS. *Sociological Methods & Research, 47*(4), 687-728. http://doi.org/10.1177/0049124117729691

Murat, M., & Frederic, P. (2015). Institutions, culture and background: the school performance of immigrant students. *Education Economics, 23*(5), 612-630. https://doi.org/10.1080/09645292.2014.894497

Muthén, B., & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups*. http://www.statmodel.com/download/PolAn.pdf

Muthén, B., & Asparouhov, T. (2014). IRT studies of many groups: the alignment method. *Frontiers in Psychology, 5*, 1- 7. http://doi.org/10.3389/fpsyg.2014.00978

Muthén, B., & Asparouhov, T. (2018). Recent methods for the study of measurement invariance with many groups: alignment and random effects. *Sociological Methods & Research, 47*(4), 637-664. http://doi.org/10.1177/0049124117701488

Muthén, B., Khoo, S., & Gustafsson, J. (1997). *Multilevel latent variable modeling in multiple populations*. Graduate School of Education & Information Studies. University of California, Los Angeles. Retrieved from https://s3.amazonaws.com/academia.edu.documents/41460205/Article_074.pdf?response-content-disposition=inline%3B%20filename%3DMultilevel_Latent_Variable_Modeling_in_M.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIWOWYYGZ2Y53UL3A%2F20200217%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Date=20200217T011009Z&X-Amz-Expires=3600&X-Amz-SignedHeaders=host&X-Amz-Signature=2137dfedf50aed812c385b0e71897dc5bd4cf8b017472e5742aaebbdc7b5e2eb

Muthén, L. K., & Muthén, B. O. (1998-2011). *Mplus* (Version 8.1) [Computer software]. Muthén & Muthén. http://www.statmodel.com/

Nansel, T. R., Craig, W., Overpeck, M. D., Saluja, G., Ruan, J., & The Health Behaviour in School-aged Children Bullying Analyses Working Group. (2004). Cross-national consistency in the relationship between bullying behaviors and psychosocial adjustment. *Archives of pediatrics & adolescent medicine, 158*(8), 730-736. http://doi.org/10.1001/archpedi.158.8.730

Oishi, S. (2006). The concept of life satisfaction across cultures: an IRT analysis. *Journal of Research in Personality, 40*(4), 411-423. doi: 10.1016/j.jrp.2005.02.002

Oliveri, M. E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education, 24*(4), 349-366. http://doi.org/10.1080/08957347.2011.607063

Oliveri, M. E., Ercikan, K., & Simon, M. (2015). A framework for developing comparable multilingual assessments for minority populations: why context matters. *International Journal of Testing*, 1-20. http://doi.org/10.1080/15305058.2014.986271

Oliveri, M. E., & Lawless, R. (2018). *The validity of inferences from locally developed assessments administered globally* (Report No. ETS RR-18-35). Princeton, NJ: Educational Testing Service.

Oliveri, M. E., Olson, B. F., Ercikan, K., & Zumbo, B. D. (2012). Methodologies for investigation item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing, 12*(3), 203-223. http://doi.org/10.1080/15305058.2011.617475

Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international large-scale assessment claims and country aims* (Report No. ETS RR-18-27). Princeton, NJ: Educational Testing Service.

Oliveri, M. E., & von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: implications for the test and sampling designs. *International Journal of Testing, 16(*3), 220-239. http://doi.org/10.1080/15305058.2015.1069743

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling, 53*(3), 315-333. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Oliveri, M. E., & von Davier, M. (2013). Toward increasing fairness in score scale calibrations employed in international large scale assessments. *International Journal of Testing, 14*(1), 1-21. http://doi.org/10.1080/15305058.2013.825265

Organisation for Economic Co-operation and Development [OECD] (2015). *Immigrant students at school: easing the journey towards integration.* https://dx.doi.org/10.1787/9789264249509-en

Organisation for Economic Co-operation and Development [OECD]. (2016). *Country note. Key findings from PISA 2015 for the United States*. https://www.oecd.org/pisa/PISA-2015-United-States.pdf

Organisation for Economic Co-operation and Development [OECD]. (2017). *PISA 2015 results (volume III): students' well-being*. http://doi.org/10.1787/9789264273856-en

Organisation for Economic Co-operation and Development [OECD]. (2018a). *PISA 2018 Technical Report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/#d.en.423800

Organisation for Economic Co-operation and Development [OECD]. (2018b). *Sampling in PISA*. https://www.oecd.org/pisa/pisaproducts/SAMPLING-IN-PISA.pdf

Organisation for Economic Co-operation and Development [OECD]. (2019a). *PISA 2018 Assessment and Analytical Framework*. PISA, OECD Publishing. https://doi.org/10.1787/b25efab8-en

Organisation for Economic Co-operation and Development [OECD]. (2019b). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. PISA, OECD Publishing. https://doi.org/10.1787/acd78851-en.

Pendergast, L. L., von der Embse, N., Kilgus, S. P., & Eklund, K. R. (2017). Measurement equivalence: a non-technical primer on categorical multi-group confirmatory factor analysis in school Psychology. *Journal of School Psychology, 60,* 65-82. https://doi.org/10.1016/j.jsp.2016.11.002

Pivovarova, M., & Powers, J. M. (2019). Generational status, immigrant concentration and academic achievement: comparing first and second-generation immigrants with third-plus generation students. *Large-scale Assessments in Education, 7*, 1-18. http://doi.org/10.1186/s40536-019-0075-4

Powers, J. M., & Pivovarova, M. (2017). Analyzing the achievement and isolation of immigrant and U.S.-born students: insights from PISA 2012. *Educational Policy, 31*(6), 830-857. http://doi.org/10.1177/0895904817719530

Ratha, Dilip K.; De, Supriyo; Schuettler, Kirsten; Seshan, Ganesh Kumar; Yameogo, Nadege Desiree. (2018). *Migration and Remittances: Recent Developments and Outlook - Transit Migration (English)* (Migration and development brief no. 29). World Bank Group. http://documents.worldbank.org/curated/en/907921534404019026/Migration-and-remittances-recent-developments-and-outlook-transit-migration

Rikoon, S. H., & Midkiff, B. (2018). *Using the SuccessNavigator ® Assessment to assess change over time: a longitudinal measurement invariance study* (ETS Research Report No. RR-18-29). Retrieved from https://onlinelibrary.wiley.com/doi/epdf/10.1002/ets2.12216

Roberson, N. D., & Zumbo, B. D. (2019). Migration background in PISA's measure of social belonging: using a diffractive lens to interpret multi-method DIF studies. *International Journal of Testing*, *19*(4), 363-389. http://doi.org/10.1080/15305058.2019.1632316

Rosen, L. H., Beron, K. J., & Underwood, M. K. (2013). Assessing peer victimization across adolescence: measurement invariance and developmental change. *Psychological Assessment, 25*(1), 1-11. http://doi.org/10.1037/a0028985

RStudio Team (2020). *RStudio: Integrated Development for R* (Version 1.3.1093) [Computer software]. http://www.rstudio.com/

Rubinstein-Avila, E. (2016). Immigrant and refugee students across "receiving" nations: to what extent can educators rely on PISA for answers? *The Clearing House, 89*(3), 79-84. http://doi.org/10.1080/00098655.2016.1168350

Rutkowski, L., & Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: a look back and a way forward. *Scandinavian Journal of Educational Research, 62*(3), 354-367. http://doi.org/10.1080/00313831.2016.1261044

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31-57. http://doi.org/10.1177/0013164413498257

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39-51. http://doi.org/10.1080/08957347.2016.1243540

Samara, M., Foody, M., Gobel, K., Altawil, M., & Scheithauer, H. (2019). Do cross-national and ethnic group bullying comparisons represent reality? Testing instruments for structural equivalence and structural isomorphism. *Frontiers in Psychology, 10*(1621), 1-14. https://doi.org/10.3389/fpsyg.2019.01621

Sandilands, D., Oliveri, M. E., Zumbo, B. D., & Ercikan, K. (2013). Investigating sources of differential item functioning in international large-scale assessments using a confirmatory approach. *International Journal of Testing, 13*(2), 152-174. http://doi.org/10.1080/15305058.2012.690140

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research, 27*(7), 1745-1755. http://doi.org/10.1007/s11136-017-1680-8

Scherer, R., Nilsen, T., & Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology, 7*(10), 1-16. http://doi.org/10.3389/fpsyg.2016.00110

Schlagel, C., & Sarstedt, M. (2016). Assessing the measurement invariance of the four-dimensional cultural intelligence scale across countries: a composite model approach. *European Management Journal, 34*, 633-649. http://doi.org/10.1016/j.emj.2016.06.002

Seddig, D., & Lomazzi, V. (2019). Using cultural and structural indicators to explain measurement noninvariance in gender role attitudes with multilevel structural equation modeling. *Social Science Research, 84*, 1-14. http://doi.org/10.1016/j.ssresearch.2019.102328

Sideridis, G. D., Tsaousis, I., & Al-Sadaawi, A. (2018). Assessing construct validity in math achievement: an application of multilevel structural equation modeling (MSEM). *Frontiers in Psychology, 9*(1451), 1-17. doi: 10.3389/fpsyg.2018.01451

Sireci, S. G. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in Psychology* (pp. 216-243). New York, NY: Cambridge University Press.

Sireci, S. G. (2015). Beyond ranking of nations: innovative research on PISA. *Teachers College Record, 117*, 1-8. Retrieved from https://www.tcrecord.org/search.asp?kw=sireci&x=0&y=0

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P.F. Merenda & C. D. Spielberg (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-115). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Spees, L. P., Potochnick, S., & Perreira, K. M. (2016). The academic achievement of limited English proficient (LEP) youth in new and established immigrant states: lessons from the National Assessment of Educational Progress (NAEP). *Education Policy Analysis Archives, 24*(99), 1-31. http://doi.org/10.14507/epaa.24.2130

Svetina, D., Rutkowski, L.., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: an illustration using Mplus and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal, 27*(1), 111-130. http://doi.org/10.1080/10705511.2019.1602776

Spielberger, C. D., Moscoso, M. S., & Brunner, T. M. (2005). Cross-cultural assessment of emotional states and personality traits. In R. K. Hambleton, P.F. Merenda & C. D. Spielberg (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 343-364). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Teltemann, J., & Schunck, R. (2016). Education systems, school segregation, and second-generation immigrants' educational success: evidence from a country-fixed effects approach using three waves of PISA. *International Journal of Comparative Sociology, 57*(6), 401-424. http://doi.org/10.1177/0020715216687348

United Nations. Department of Economic and Social Affairs (2017). *International Migration Report 2017: Highlights*. https://www.un.org/en/development/desa/population/migration/publications/migrationreport/docs/MigrationReport2017_Highlights.pdf

United Nations Educational, Scientific and Cultural Organization [UNESCO]. (2019). *Behind the numbers: ending school violence and bullying*. https://unesdoc.unesco.org/ark:/48223/pf0000366483

van de Vijver, F. J. R., Avvisati, F., Davidov, E., Eid, M., Fox, J. P., Le Donné, N.,…van de Schoot, R. (2019). *Invariance analyses in large-scale studies* (Working Papers No. 201). Retrieved from https://www.oecd-ilibrary.org/docserver/254738dd-

en.pdf?expires=1573337653&id=id&accname=guest&checksum=C31A91757FB1D3E443B5C2DE875DA2EC

van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: a review of concepts, models, and data analytic procedures. In D. Matsumoto & F. J. R. van de Vijver (Eds.), *Cross-cultural research methods in Psychology* (pp. 17-45). New York, NY: Cambridge University Press.

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P.F. Merenda & C. D. Spielberg (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-64). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

van Dijk, W., Gage, N. A., & Grasley-Boy, N. (2019). The relation between classroom management and mathematics achievement: a multilevel structural equation model. *Psychology in the Schools*, *56*, 1173-1186. http://doi.org/10.1002/pits.22254

van Hemert, D. A., Poortinga, Y. H., & van de Vijver, F. J. R. (2007). Emotion and culture: a meta-analysis. *Cognition and Emotion*, *21*(5), 913-943. http://doi.org/10.1080/02699930701339293

Verdín, D., & Godwin, A. (2017, October). *Testing for measurement invariance in Engineering identity constructs for first-generation college students* [Paper presentation]. 2017 IEEE Frontiers in Education Conference (FIE), Indianapolis, IN, United States.

Vessey, R., Strout, T. D., DiFazio, R. L., & Walker, A. (2014). Measuring the youth bullying experience: a systematic review of the psychometric properties of available instruments. *Journal of School Health*, *84*(12), 819-843. http://doi.org/10.1111/josh.12210

Vigil-Colet, A., Ruiz-Pamies, M., Anguiano-Carrasco, C., & Lorenzo-Seva, U. (2012). The impact of social desirability on psychometric measures of aggression. *Psicothema*, *24*(2), 310-315. http://www.psicothema.com/pdf/4016.pdf

Volante, L., Klinger, D., Bilgili, O., & Siegel, M. (2017). Making sense of the performance (dis)advantage for immigrant students across Canada. *Canadian Journal of Education*, *40*(3), 329-361. http://doi.org/ 10.2307/90014781

Volk, A. A., Veenstra, R., & Espelage, D. L. (2017). So you want to study bullying? Recommendations to enhance the validity, transparency, and comparability of bullying research. *Aggression and Violent Behavior*, *36*, 34-43. http://doi.org/10.1016/j.avb.2017.07.003

Weijters, B., Baumgartner, H., & Geuens, M. (2016). The calibrated sigma method: an efficient remedy for between-group differences in response category use on Likert

scales. *International Journal of Research in Marketing, 33*(4), 944-960. https://doi.org/10.1016/j.ijresmar.2016.05.003

Wendt, H., Kasper, D., & Trendtel, M. (2017). Assuming measurement invariance of background indicators in international comparative educational achievement studies: a challenge for the interpretation of achievement differences. *Large-scale Assessments in Education, 5*, 1-34. http://doi.org/10.1186/s40536-017-0043-9

Williams, B. D., Chandola, T., & Pendleton, N. (2018). An application of Bayesian measurement invariance to modeling cognition over time in the English longitudinal study of ageing. *International Journal of Methods in Psychiatric Research, 27*(4), 1-8. http://doi.org/10.1002/mpr.1749

Wolgast, A., & Donat, M. (2019). Cultural mindset and bullying experiences: an eight-year trend study of adolescents' risk behaviors, internalizing problems, talking to friends, and social support. *Children and Youth Services Review, 99*, 257-269. http://doi.org/10.1016/j.childyouth.2019.02.014

Wu, J., Lin, J. J. H., Nian, M., & Hsiao, Y. (2017). A solution to modeling multilevel confirmatory factor analysis with data obtained from complex survey sampling to avoid conflated parameter estimates. *Frontiers in Psychology, 8*(1464), 1-19. http://doi.org/10.3389/fpsyg.2017.01464

Zigler, C. K., & Ye, F. (2019). A comparison of multilevel mediation modeling methods: recommendations for applied researchers. *Multivariate Behavioral Research, 54*(3), 338-359. http://doi.org/10.1080/00273171.2018.1527676

Zyphur, M. J., Zhang, Z., Preacher, K. J., & Bird, L. J. (2019). Moderated mediation in multilevel structural equation models: decomposing effects of race on math achievement within versus between high schools in the United States. In S. E. Humphrey & J. M. LeBreton (Eds.), *The Handbook of Multilevel Theory, Measurement, and Analysis* (pp. 473- 494). American Psychological Association. http://dx.doi.org/10.1037/0000115-021